ABSTRACT

Title of Dissertation: SPATIAL ANALYSIS AND OPTIMIZATION MODELS FOR BETTER UNDERSTANDING OF ILLICIT DRUG SUPPLY CHAINS WITH SPARSE AND INCOMPLETE DATA

Eunseong Jang
Doctor of Philosophy, 2026

Dissertation Directed by: Associate Professor, Margrét V. Bjarnadóttir
Decision, Operations & Information Technologies

Professor, S. Raghavan,
Decision, Operations & Information Technologies

Illicit drug trafficking networks are difficult to observe directly due to their clandestine nature, resulting in sparse, incomplete, and often inconsistent data. These limitations hinder law enforcement efforts, including identifying trafficking routes, detecting emerging hot spots, and understanding the underlying structure of illicit supply chains. This dissertation develops methodological approaches to improve the analysis and inference of cocaine trafficking networks under such data constraints. Using cocaine-related data from Colombia and the United States, the research applies geographically weighted regression analysis, introduces an improved spatial data analysis method, and develops a data-driven optimization framework to generate actionable insights for law enforcement and policymakers.

The first study investigates region-specific drivers of cocaine trafficking in Colombia using a local geographically weighted regression framework, addressing spatial heterogeneity, sparse observations,

and separation issues. The results reveal substantial geographic variation in the factors associated with cocaine trafficking destinations, demonstrating that the determinants of trafficking activity differ considerably across regions.

The second study examines the use of Local Indicators of Spatial Association (LISA) for identifying trafficking hot spots. Using crack cocaine seizure data from the United States, the study shows that conventional LISA methods may overlook meaningful spatial patterns in zero-heavy datasets. An enhanced LISA framework is proposed to better capture local clustering under extreme sparsity, resulting in the identification of hot spot patterns that differ meaningfully from those obtained using traditional approaches.

The third study develops a data-driven optimization model to infer hidden cocaine flow networks in the United States from incomplete and noisy observations, including seizure records, undercover purchases, and overdose data. The proposed framework constructs plausible trafficking networks consistent with observed data and allows the exploration of alternative network structures through adjustable design parameters. Validation against anecdotal trafficking intelligence demonstrates the capability of the approach for identifying potential supply chain structures.

Collectively, these studies contribute new methodologies for analyzing illicit supply chains in settings characterized by limited information, spatial heterogeneity, and partially observed network structures. The findings demonstrate how statistical and optimization-based approaches can enhance understanding of cocaine trafficking systems and support more effective interdiction and enforcement strategies.

SPATIAL ANALYSIS AND OPTIMIZATION MODELS FOR BETTER UNDERSTANDING OF ILLICIT DRUG SUPPLY CHAINS WITH SPARSE AND INCOMPLETE DATA

by

Eunseong Jang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2026

Advisory Committee:
Professor Margrét V. Bjarnadóttir, Chair/Advisor
Professor S. Raghavan, Co-Chair/Co-Advisor
Dr. Marcus A. Boyd
Professor Catherine Nakalembe
Professor Greg Midgette

© Copyright by
Eunseong Jang
2026

Acknowledgments

As I reflect on my five-year journey as a Ph.D. student at the Robert H. Smith School of Business, I am struck by how quickly the time has passed and how much I have grown throughout this experience. I would not have reached this point without the guidance, encouragement, and support of many people.

First and foremost, I would like to express my deepest gratitude to my advisors, Professor Margrét V. Bjarnadóttir and Professor S. Raghavan. Their mentorship, patience, and unwavering support have shaped both my research and my development as a scholar. I am sincerely grateful for the opportunities they provided and for their confidence in me throughout this journey.

I would also like to thank Dr. Marcus A. Boyd for helping me develop a deeper understanding of geographic analysis and for providing valuable insights that strengthened this dissertation. I am grateful to my committee members, Professor Catherine Nakalembe and Professor Greg Midgette, for their thoughtful feedback, guidance, and constructive suggestions.

My sincere thanks also go to Professor Galia J. Benítez of Michigan State University for her assistance in constructing the Colombian cocaine trafficking dataset and for generously sharing her knowledge of Colombia's local and geopolitical context. Her expertise was invaluable to this research.

I am deeply appreciative of the support and opportunities provided by the Robert H. Smith School of Business throughout my doctoral studies. It has been a privilege to learn and grow within such a supportive academic community.

Finally, I would like to thank my family for their unconditional love, encouragement, and unwavering support. Their belief in me sustained me through the challenges of this journey.

Table of Contents

Acknowledgements ii

Table of Contents iii

List of Tables v

List of Figures vi

Abbreviations vii

Chapter 1: Introduction 1

Chapter 2: Identifying Region-Specific Drivers of Cocaine Trafficking in Colombia Using Local Geographically Weighted Regression 4

2.1 Literature Review . . . 6
2.1.1 Geographically Weighted Regression . . . 6
2.1.2 Coca Cultivation and Cocaine Trafficking in Colombia . . . 8
2.2 Data Overview . . . 10
2.2.1 Dependent Variable . . . 10
2.2.2 Independent Variables . . . 12
2.2.3 Final Dataset . . . 17
2.3 Methodology . . . 20
2.3.1 Local Geographically Weighted Regression with Penalized Maximum Likelihood Estimation . . . 20
2.4 Results and Discussion . . . 25
2.4.1 Spatial Variation of Regression Coefficients . . . 25
2.4.2 Case Study: San Antero, Córdoba, and Convención, Norte de Santander . . . 40
2.4.3 Case Study: Potential Cocaine Trafficking Destinations Based on the Local GWR PML Model . . . 43
2.4.4 Sensitivity Analysis . . . 45
2.4.5 Comparisons to Another Approach: GWR LASSO . . . 48
2.5 Conclusion . . . 50

Chapter 3: Improved LISA Analysis for Zero-Heavy Crack Cocaine Seizure Data to Detect Trafficking Hot Spots 53

3.1 Literature Review . . . 54
3.2 Data Overview . . . 56
3.3 Local Indicator of Spatial Association (LISA) Statistic: Local Moran's I . . . 59
3.4 Statistical Significance Test of LISA Statistics: Permutation Inference . . . 60
3.5 Limitations in LISA Analysis with the Standard Method . . . 61
3.6 Improved LISA Analysis . . . 65
3.6.1 Permuted i Method . . . 65

3.6.2 Moderate Label and Combined Methods 66
3.7 Case Study with the Northeastern States and Florida 68
3.8 Conclusion 71

Chapter 4: Data-Driven Optimization Model for Cocaine Flow Networks Inference 73
4.1 Literature Review 73
4.2 Data Overview 75
4.3 Network Optimization Framework 76
4.3.1 Model 1: Relaxed Seizure Constraint 77
4.3.2 Model 2: Relaxed Seizure and Price Directionality 81
4.3.3 Model 3: Statistical Price Penalty Relaxation 83
4.4 Conclusion 88

Chapter 5: Conclusion 91

Appendix A: Geographical Variation of Aggregated Data: Remaining Variables 93

Appendix B: Firth's Penalized Likelihood Method 96

Bibliography 99

List of Tables

2.1 Variables used in the analysis. 18
2.2 An example of quasi-complete separation in the dependent variable (Cocaine Destination), with the binary indicator for airport's presence in a *municipio* (Airport). 22
2.3 Summary of regression coefficients obtained with the local GWR PML and the global model. 27
2.4 Local GWR PML model results of San Antero, Córdoba, and Convención, Norte de Santander. 40
2.5 Local GWR PML model results of Medio San Juan, Chocó, and Cumaral, Meta. 43
2.6 Sensitivity analysis for n_y and n_{drop} for each variable. 47

3.1 Notation for Section 3.3 and 3.4. 61

4.1 Decision variables and key parameters. 76

List of Figures

2.1 Cocaine hydrochloride destinations in 2016 and 2017 highlighted in red. 11
2.2 Geographical variation of (a) coca_area (log of ha), (b) seizures (log of kg), (c) lab_reported, and (d) lab_prob data. 19
2.3 Flowchart for local GWR PML model fitting 23
2.4 Violin box plot of estimated regression coefficients using the local GWR PML model. . 26
2.5 Estimated regression coefficients of (a) coca_area, (b) seizures, and (c) lab_prob using the local GWR PML model. 30
2.6 Estimated regression coefficients of (a) left_wing, (b) right_paramilitary, and (c) left_right using the local GWR PML model. 33
2.7 Estimated regression coefficients of (a) airport, (b) ferry, (c) military, and (d) police using the local GWR PML model. 36
2.8 Estimated regression coefficients of (a) river_length, (b) road_length, and (c) population using the local GWR PML model. 38
2.9 Natural regions of Colombia (Milenioscuro 2014). 39
2.10 AUC of the local GWR PML model obtained with various n_y and n_{drop}. 46
2.11 Estimated regression coefficients of (a) coca_area, (b) seizures, and (c) lab_prob using local GWR lasso model. 49

3.1 The number of crack cocaine seizures in January 2020 by county. 57
3.2 Histogram of county-level crack cocaine seizure counts in January 2020. 58
3.3 Moran's scatter plots of (a) centered crack cocaine seizure counts ($z_i = x_i - \bar{x}$) in January 2020 versus the sum of its neighbors' seizure counts $\left(\sum_{j=1}^{N} w_{ij} z_j\right)$ and (b) its significance region with $M = 999$, $k = 5$, and $\alpha = 0.1$. 62
3.4 Moran's scatter plots of (a) centered values z_i of normally distributed random samples $x_i \sim N(10, 25)$ with sample size of 1,518 and (b) its significance region with $M = 999$, $k = 5$, and $\alpha = 0.1$. 63
3.5 Significance regions using the permuted i approach for centered crack cocaine seizure counts z_i vs. sum of neighbors' seizure counts $\left(\sum_{j=1}^{N} w_{ij} z_j\right)$ with $M = 999$, $k = 5$, and $\alpha = 0.1$. 66
3.6 Significance region with the zero-heavy data using the moderate label with (a) the standard approach and (b) the permuted i approach with $M = 999$, $k = 5$, and $\alpha = 0.1$. 67
3.7 The Northeastern states' LISA labels for crack cocaine seizure data from January 2020 under (a) standard, (b) permuted i, (c) moderate label, and (d) combined. 69
3.8 Florida's LISA labels for crack cocaine seizure data in January 2020 under (a) standard and (b) permuted i approaches. 71

4.1 The optimal network from Model 1 with $\alpha = 0.9$ and $\varepsilon = 0.4$. 80
4.2 Routes for cocaine flows across the U.S. 80

4.3 The optimal network from Model 2 with $\alpha = 0.9$ and $\varepsilon = 1.0$. 82
4.4 The optimal network for Model 3 with two different parameter settings (a) $\alpha = 1$ and $\beta = 5$, (b) $\alpha = 1$ and $\beta = 10$. 87

A.1 Geographical variation of (a) left_wing and (b) right_paramilitary data. 93
A.2 Geographical variation of (a) airport, (b) ferry, (c) military, and (d) police data. 94
A.3 Geographical variation of (a) river_length, (b) road_length, and (c) population data. . . 95

List of Abbreviations

AUC	Area Under the Curve
CDC	Centers for Disease Control and Prevention
cdf	cumulative distribution function
DEA	Drug Enforcement Administration
ELN	Ejército de Liberación Nacional (National Liberation Army of Colombia)
FARC	Fuerzas Armadas Revolucionarias de Colombia (Revolutionary Armed Forces of Colombia)
GWR	Geographically Weighted Regression
HIDTA	High Intensity Drug Trafficking Areas
IMR	Inverse Mills Ratio
INDEPAZ	Instituto de Estudios para el Desarrollo y la Paz (Institute of Studies for Development and Peace)
LASSO	Least Absolute Shrinkage and Selection Operator
LISA	Local Indicator of Spatial Association
LLM	Large Language Model
MGWR	Multiscale Geographically Weighted Regression
MIP	Mixed-Integer Programming
MLE	Maximum Likelihood Estimation
ODC	Observatorio de Drogas de Colombia Drug Observatory of Colombia
OSM	OpenStreetMap
pdf	probability distribution function
PML	Penalized Maximum Likelihood
SIMCI	Sistema Integrado de Monitoreo de Cultivos Ilcitos (Integrated Monitoring System for Illicit Crops)
STRIDE	System to Retrieve Information on Drug Evidence
UNODC	United Nations Office on Drugs and Crime

Chapter 1: Introduction

Illicit activities, including drug trafficking, money laundering, and environmental crime, are growing global threats to public security and the rule of law. Among them, illegal drug trafficking alone generates an estimated $870 billion annually (Bjarnadottir et al. 2024). Among various types of illicit drugs, cocaine was the second most seized illicit drug worldwide in 2024 after cannabis (UNODC 2024). The 2024 National Drug Threat Assessment (Drug Enforcement Administration 2024) reports that cocaine consumption levels in the U.S. remain high despite increases in cocaine seizures, signaling the need for more effective enforcement strategies. This underscores the need for more effective enforcement strategies to enhance cocaine seizure operations. Efficiently disrupting cocaine supply chains begins with accurate predictions of trafficking routes and a better understanding of the hidden drug supply chain network.

The vast majority of the world's coca cultivation and cocaine production occurs in the Andean region, particularly Colombia, Peru, and Bolivia (UNODC 2024). Coca leaves cultivated in remote rural areas are processed into coca paste and subsequently refined into cocaine hydrochloride in clandestine laboratories (UNODC 2021). From these production zones, cocaine is transported through complex networks of rivers, roads, air routes, maritime corridors, and transshipment hubs before reaching wholesale and retail markets in North America, Europe, and beyond. The supply chain continuously adapts to enforcement pressure by shifting cultivation sites (Ruiz et al. 2016), trafficking routes, and transportation methods, making it difficult for authorities to identify critical nodes and effectively allocate interdiction resources. As a result, understanding the structure and dynamics of cocaine trafficking networks remains a central challenge for policymakers and law enforcement agencies seeking to reduce cocaine availability.

Accurately mapping such network supply chains is challenging due to the clandestine nature

of illicit activities. Unlike legal supply chains, trafficking organizations deliberately conceal their operations, resulting in limited and fragmented observations of production, transportation, and distribution activities. Available data, such as seizures, prices, arrests, laboratory detections, and anecdotal intelligence reports, are often sparse, incomplete, and subject to substantial measurement error and reporting bias (Bjarnadottir et al. 2024). An additional challenge, critical to the use of such data, is that limited or no ground truth exists to compare the prediction results as a reference.

Motivated to overcome these challenges, this dissertation seeks to enhance the understanding cocaine trafficking networks with the ultimate goal of network disruption through three studies: (i) an analysis of trafficking patterns in Colombia using a geographically weighted regression model to uncover region-specific drivers, (ii) an improved spatial analysis method for detecting trafficking hot spots with zero-heavy data, and (iii) a data-driven optimization framework to infer cocaine flow networks across the United States.

Each study addresses a critical gap in conventional methods when leveraging real-world, sparse, and often incomplete data to generate insights for drug policy and enforcement. The first study develops a localized regression framework tailored to partially observed, sparse, and spatially heterogeneous data on cocaine trafficking in Colombia. The results reveal regional variation in the factors associated with cocaine trafficking destinations. For example, coca cultivation is an important factor along the Pacific coast, whereas river and road accessibility are the important drivers in more central inland areas. Along the Caribbean coast, recorded trafficking activity is strongly associated with the presence of police stations, while airports consistently emerge as a factor predictive of cocaine destinations in northeastern Colombia near the Venezuelan border.

Second, we present an enhanced spatial exploratory data analysis technique that improves the identification of local trafficking clusters in areas that would otherwise be overlooked. The proposed

method addresses the limitations of traditional spatial cluster detection when applied to zero-heavy data. Its application to U.S. cocaine seizure data provides a broader interpretation of spatial seizure patterns and identifies additional trafficking hot spots that the conventional method fails to detect.

Finally, in the third study, we develop a novel optimization approach to combine multiple, incomplete, and inconsistent data with the goal of constructing the unknown cocaine supply chain network in the United States. Using secondary trafficking data, the model identifies plausible state-level cocaine flows without relying on prior knowledge of the network's structure. By addressing uncertainty and inconsistency across the data, the framework reveals previously unidentified cocaine flows originating near the northern U.S. border. We further generate a set of evidence-based scenarios of cocaine flows across different model formulations and parameter values, enabling law enforcement agencies and policymakers to evaluate competing hypotheses about cocaine trafficking activity.

Chapter 2: Identifying Region-Specific Drivers of Cocaine Trafficking in Colombia Using Local Geographically Weighted Regression

The first study of my thesis develops a modeling framework that utilizes a localized regression to investigate the key drivers of cocaine trafficking in Colombia, when there is available information on cocaine trafficking routes. The objective is to uncover meaningful insights about cocaine transportation in Colombia by integrating cocaine trafficking records, information on coca cultivation, interdiction activity, and geographic and demographic data. Specifically, we seek to identify key factors that make certain locations more or less likely to be cocaine trafficking destinations. We focus on destinations of cocaine hydrochloride, the final refined form of cocaine intended for consumer markets, to capture the movement of consumer-ready cocaine within Colombia and heading toward international export routes. The proposed framework can be applied to any stage of cocaine production, and generalizes for other illicit drugs.

A key concern when mapping Colombia's cocaine supply chain is the incomplete information on cocaine destinations and production sites (laboratories). The available data only provide a partial picture of the underlying trafficking network, which complicates statistical inference and model estimation. Compounding this challenge is the geographic heterogeneity in determinants of trafficking activity. The effects of socioeconomic, geographic, and enforcement-related factors on trafficking patterns are unlikely to be spatially uniform across Colombia. This heterogeneity is reflected in our data. For example, coca cultivation is concentrated primarily in southwestern Colombia, whereas cocaine refinement more commonly takes place in central and northern regions before domestic and international distribution. In addition, cocaine trafficking networks are not governed by a single homogeneous

organization (Kenney 2007). Consequently, the relationships between independent variables and trafficking destinations are likely to vary substantially across Colombia depending on region-specific characteristics. As a result, applying a single global regression model would obscure important regional differences underlying trafficking dynamics.

To account for this spatial heterogeneity, we adopt a geographically weighted regression (GWR) (Brunsdon et al. 1996). More specifically, we develop and apply a localized logistic regression framework addressing technical challenges arising from data sparsity. We incorporate Firth's penalized maximum likelihood (PML) estimation (Firth 1993), together with neighbor-selection rules designed to mitigate small-sample issues as we explain in Section 2.3.1.1. We show that the proposed local approach yields stable and reliable coefficient estimates while preserving the ability to capture meaningful spatial heterogeneity in trafficking-related patterns.

Our analysis is conducted at the *municipio* and *corregimiento* levels. A *municipio* is an administrative unit equivalent to a county in the United States (U.S.)[1]. *Corregimientos* are usually less populated than a *municipio*, but at the same administrative level. For simplicity, we hereafter refer to both *municipios* and *corregimientos* collectively as *municipios*. In total, mainland Colombia contains 1,120 *municipios*. Unlike global regression approaches, GWR estimates a separate local regression model for each location using observations from neighboring *municipios*. This localized estimation enables us to capture spatially varying relationships between different factors and estimate the probability of a *municipio* being a cocaine trafficking destination.

Beyond its application to cocaine trafficking, the proposed framework contributes to the analysis of illicit activities or rare-event phenomena characterized by spatial heterogeneity, sparse observations, and partially observed binary features. By combining geographically weighted modeling with bias-

[1]The primary administrative division in Colombia is departamento (department), which is similar to a U.S. state.

reducing estimation techniques, the proposed approach offers a flexible methodological framework for studying complex criminal and clandestine activities in settings where conventional global or penalized regression approaches may produce unstable or misleading results.

2.1 Literature Review

2.1.1 Geographically Weighted Regression

The concept of geographically weighted regression was originally introduced by Brunsdon et al. (1996). Unlike conventional global regression approaches, GWR estimates separate local regression models for each geographic location while assigning weights to neighboring observations according to spatial proximity. These weights may be binary or continuous functions of distance, typically decreasing as observations are farther from the focal location (Brunsdon et al. 1996). The size of the local neighborhood can be fixed or optimized based on model selection criteria such as the Akaike Information Criterion (AIC) (Comber et al. 2018). By allowing regression coefficients to vary spatially, GWR can capture geographic heterogeneity in the relationships between independent and dependent variables. Owing to this flexibility, GWR has been widely applied in studies of ecosystems (Wang et al. 2005), environmental processes (Atkinson et al. 2003), and transportation systems (He et al. 2021).

One of the primary methodological challenges in GWR is multicollinearity among independent variables. Wheeler and Tiefelsdorf (2005) emphasized that multicollinearity can become particularly severe in GWR because local regressions are estimated using substantially smaller samples than global models. This may produce highly correlated coefficient estimates, inflated standard errors, and unstable inference. To overcome these issues, several extensions of GWR incorporating penalized regression techniques have been proposed. For example, ridge-based GWR models were introduced by

Wheeler (2007) and Bárcena et al. (2014), while geographically weighted Least Absolute Shrinkage and Selection Operator (LASSO) models were proposed by Wheeler (2009) and Wang and Li (2017). Bárcena et al. (2014) further proposed an adaptive sampling strategy that enlarges local sample sizes when the condition number[2] of an independent variable exceeds a specified threshold, thereby reducing the effects of multicollinearity. This strategy differs from our approach that focuses on the heterogeneity of the independent variables.

Subsequent research extended these approaches to logistic GWR settings involving binary dependent variables. For example, Yoneoka et al. (2016) proposed a geographically weighted logistic LASSO model, while Comber and Harris (2018) introduced an elastic-net extension combining ridge and LASSO penalties. Although these methods address multicollinearity and improve local estimation stability, they do not explicitly address the problem of (quasi-)complete separation (Heinze and Schemper 2002) in GWR models. Separation occurs when one or more independent variables perfectly or nearly perfectly distinguish the binary outcome, often resulting in inflated or unstable coefficient estimates. This issue is especially problematic in GWR because local estimation frequently involves small and highly imbalanced samples. Heinze and Schemper (2002) demonstrated that Firth's penalized maximum likelihood (PML) (Firth 1993) method provides a robust solution to the separation issue in logistic regression.

Fotheringham and colleagues introduced a multiscale geographically weighted regression (MGWR; Fotheringham et al. 2017), which allows each independent variable to have its own optimal bandwidth. MGWR relaxes the assumption that all independent variables operate at the same spatial scale and estimates coefficients via a generalized additive modeling framework in which independent variables are iteratively added through regressions on model residuals. However, MGWR poses practical challenges

[2]The ratio of the largest to the smallest singular value of a covariate matrix (Bárcena et al. 2014).

in our setting. In particular, extending MGWR to logistic regression is nontrivial, as the use of residuals in the additive model is less straightforward for binary outcomes. Moreover, the sparsity of our data increases the risk of overfitting when variable-specific bandwidths are estimated. For these reasons, we apply GWR in our setting, which applies a common bandwidth across all independent variables.

2.1.2 Coca Cultivation and Cocaine Trafficking in Colombia

There is a growing literature focusing on the effects of eradication policies on coca cultivation dynamics in Colombia. Reyes (2014) found that coca eradication increased subsequent coca cultivation. Ruiz et al. (2016) reported that eradication reduced overall coca production while displacing the coca cultivation areas. In contrast, Dávalos and Morales (2023) provided quantitative evidence that aerial fumigation affects coca cultivation beyond the targeted *municipio*.

A study by Grajales-Marín et al. (2025) employed GWR to identify spatial non-stationarity in the effects of socioeconomic variables on coca cultivation within the Nariño *departamento* of Colombia. However, their analysis was limited to a single department and focused primarily on coca cultivation rather than trafficking dynamics. From a different perspective, Andrade et al. (2024) examined the illicit drug trafficking patterns in Colombia's maritime space, providing insights into trafficking routes beyond the mainland. In contrast, this chapter focuses on spatial variation in the drivers of cocaine trafficking destinations across Colombian *municipios* and the diverse geographic, socioeconomic, enforcement, and conflict-related factors associated with them.

Other studies have focused on the behavioral and environmental dimensions of the cocaine supply chain. Using a field experiment, Ibanez and Martinsson (2013) demonstrated that coca farmers respond differently to anti-drug interventions depending on their perceived moral costs of cultivating coca. Mejia

and Restrepo (2016) modeled the marginal costs of subsidizing intervention strategies in Colombia to reduce a kilogram (kg) of cocaine in the retail market of the U.S, estimating that the investment in interdiction is more than three times as cost-efficient as investment in eradication. Negret et al. (2019) investigated the effects of armed conflict and coca cultivation on deforestation in Colombia, while Ruiz et al. (2013) examined the relationship between coca cultivation and environmental degradation, showing that both the drivers of coca cultivation and its association with deforestation vary substantially across regions. However, these studies generally focus on a limited set of independent factors or do not directly examine the determinants of cocaine trafficking patterns.

A single previous study focuses on data-driven discovery of cocaine trafficking networks in Colombia. Benítez et al. (2019) inferred cocaine trafficking networks by analyzing correlations in cocaine prices between pairs of *municipios* and validated the resulting network against anecdotal evidence. However, their approach relied exclusively on price information and was not designed to identify the broader set of socioeconomic, geographic, and conflict-related factors associated with cocaine trafficking.

Several studies have focused on violence in Colombia and its connection to the cocaine trade. For instance, Holmes et al. (2006) conducted a *departamento*-level analysis and found that violence committed by Colombian guerrilla and paramilitary groups was more strongly associated with economic conditions and coca-eradication efforts than with coca cultivation itself. Jiménez-García et al. (2023) documented significant relationships among violence, drug trafficking, and socioeconomic disadvantage in the city of Pereira, located in the Risaralda *departamento* of Colombia. In a qualitative study, Kenney (2007) noted that the Colombian cocaine industry is not controlled by a small number of vertically integrated cartels, but instead consists of decentralized and adaptive networks composed of numerous smaller organizations. Millán-Quijano (2020) demonstrated that homicide rates in Colombia

are associated with fluctuations in cocaine prices in major consumer markets such as the United States and Europe, while also revealing geographic heterogeneity in these relationships across *municipios*.

Collectively, these studies have provided important insights into individual mechanisms shaping illicit drug production and trafficking. However, no studies have examined a comprehensive set of socioeconomic, geographic, and enforcement-related factors and investigated how these factors relate to the spatial dynamics of cocaine trafficking in Colombia. Specifically, no previous study has conducted a nationwide *municipio*-level analysis of cocaine trafficking destinations in Colombia. Since trafficking networks are formed by a complex interplay of production, transportation opportunities, state presence, and armed-group activity, analyses focusing on a few factors or not considering spatial heterogeneity may overlook important spatial interactions among these factors.

2.2 Data Overview

This study's analysis period is 2016 and 2017, the two years immediately following the Colombian Peace Process; the negotiations between the Government of Colombia and the Fuerzas Armadas Revolucionarias de Colombia (Revolutionary Armed Forces of Colombia; henceforth FARC). Throughout the decades, FARC has been the major player in cocaine trafficking in Colombia. The Peace Process brought about substantial political and territorial instability that altered both trafficking dynamics and their detection.

2.2.1 Dependent Variable

Central information to this study is hydrochloride cocaine destinations from anecdotal evidence of cocaine trafficking routes in Colombia. The anecdotal information consists of the location names

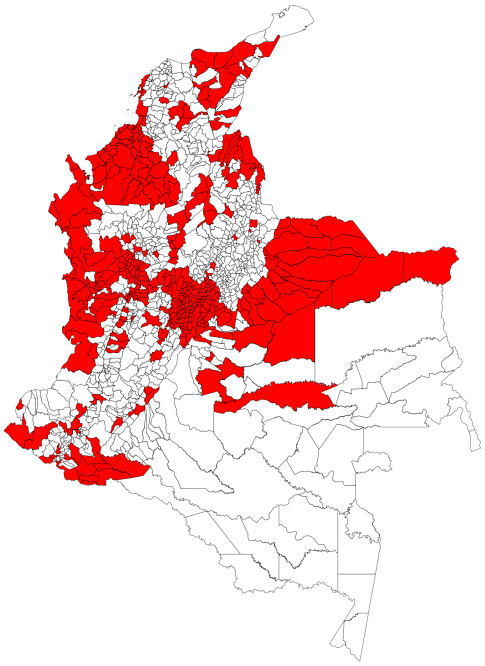


Figure 2.1: Cocaine hydrochloride destinations in 2016 and 2017 highlighted in red.

and coordinates for cocaine hydrochloride destinations in Colombia, which are confidentially reported by the Colombian Antinarcotics Police (Benítez et al. 2019). To construct the dependent variable, we aggregate the hydrochloride destinations from 2016 and 2017 at the *municipio* level. Specifically, we use a binary indicator variable equal to 1 if a *municipio* is identified as a cocaine hydrochloride destination in either year, and 0 otherwise. We note that most cocaine destinations identified in 2016 also appear in 2017. The aggregation provides a more stable representation of trafficking destinations in the analysis period, while reducing sparsity in the dependent variable. There are 384 *municipios* (approximately 34%) reported as destinations in the anecdotal data, as shown in Figure 2.1.

2.2.2 Independent Variables

We group the independent variables into three categories: cocaine-related, conflict-related, and demographic and geographic information. These categories comprise three, three, and seven variables, respectively, for a total of 14 independent variables.

2.2.2.1 Cocaine-Related Data

The first variable measures the area of coca cultivation ("coca_area" henceforth) in hectares. The data were obtained from the Sistema Integrado de Monitoreo de Cultivos Ilcitos (SIMCI, the Integrated Monitoring System for Illicit Crops) project by the United Nations Office on Drugs and Crime (UNODC). We sum up the coca-growing area for each *municipio* during the analysis period.

The next variable captures information on cocaine seizures ("seizures" henceforth) collected by *Observatorio de Drogas de Colombia* (ODC; Drug Observatory of Colombia), which is administered by the Colombian Ministry of Justice and Law (ODC 2022). The seizure data report the annual quantity of cocaine hydrochloride confiscated within each *municipio*, measured in kilograms. We use seizures measured before the analysis period rather than contemporary observations in the analysis, since interdiction activity can be endogenous to trafficking routes; drug traffickers may avoid regions with a high perceived risk of interdiction, while enforcement agencies are also more likely to allocate resources toward areas suspected of intensive trafficking activity, reflecting the interplay between drug price and enforcement (Caulkins and Reuter 1998). This simultaneous relationship can create potential back-and-forth causality between seizures and trafficking destinations. To limit the direction of causality such that only previous seizures affect the likelihood of a *municipio* being a cocaine destination, we use seizure records from 2015 as independent variables of trafficking destinations during the analysis

period.

We also address missing observations for coca cultivation or seizure data. Given the reliability of the data collection procedures employed by the ODC (2022) and UNODC (2023) for coca cultivation and seizures, we treat missing values in these variables as zeros. Specifically, since coca cultivation data is obtained by analyzing satellite images, we interpret the absence of a record for a *municipio* as no coca cultivation. Similarly, we treat missing seizure records as zero based on the assumption that all official interdiction is reported to the ODC (2022).

The laboratory dataset contains annual *municipio*-level information on cocaine-processing laboratories (cocaine hydrochloride laboratories), also collected by ODC (2022). As noted in a publication by UNODC (2021), the final product of cocaine is produced by first extracting cocaine base from coca leaves. Then, the cocaine base is chemically refined into cocaine hydrochloride through crystallization and hydrochloric acid processing in hydrochloride laboratories. The hydrochloride laboratory data are converted into a binary indicator, "lab_reported", equal to 1 if at least one cocaine hydrochloride laboratory was reported in a *municipio* during the analysis period, and 0 otherwise.

Since illicit activity is partially observed, the absence of a cocaine-processing laboratory does not necessarily imply the true absence of a production site within a *municipio*. Laboratories may be more likely to be detected and reported in areas with stronger law-enforcement presence, greater institutional capacity, or better accessibility. Consequently, observed laboratory reports may reflect not only underlying cocaine trafficking activity but also spatial differences in the detection and reporting process. Directly using reported laboratory presence may therefore lead to biased or misleading estimation. In addition, cocaine laboratories are observed only in specific regions of Colombia, as shown in Figure 2.2(c). This spatial sparsity may produce unstable coefficient estimates and complicate their interpretation.

To address these problems, we construct a laboratory probability variable (lab_prob) instead of using the binary laboratory indicator directly. We obtain the predicted probability of laboratory presence with a logistic regression model by regressing the observed laboratory indicator on coca_area, seizures, and other independent variables introduced in the following sections. Conceptually, this approach is similar to propensity-score regression adjustment (Cai et al. 2025, D'Agostino Jr 1998, Davis et al. 2021), in which the propensity score represents the conditional probability of receiving a treatment given a set of observed covariates (Rosenbaum and Rubin 1983). If laboratory presence is considered as a treatment, the predicted probability (lab_prob) serves as the corresponding propensity score. Replacing the locally sparse binary indicator with a continuous measure of expected laboratory presence, this approach can reduce estimation instability associated with the geographic concentration of reported laboratories.

2.2.2.2 Conflict Data

To examine the relationship between armed conflict and cocaine trafficking, we construct *municipio*-level binary indicators that represent the presence of armed groups (left- and right-wing) using two complementary datasets. Armed groups have long played an important role in Colombia's illicit drug economy, often exercising territorial control, facilitating trafficking activities, and engaging in violence related to the drug trade (National Drug Intelligence Center 2016, UNODC 2024). Furthermore, prior research has identified a significant association between violent crime, drug trafficking, and socioeconomic disadvantage in the city of Pereira, Risaralda, Colombia (Jiménez-García et al. 2023). We therefore expect the presence of left- or right-wing armed groups to be associated with a *municipio*'s likelihood of serving as a cocaine trafficking destination, at least for some regions.

We use two sources of data to indicate the presence of left- or right-wing armed groups. First, we use reports published by the *Instituto de Estudios para el Desarrollo y la Paz* (INDEPAZ; Institute of Studies for Development and Peace; Indepaz 2020). These reports provide information on the territorial presence of paramilitary organizations between 2008 and 2020, primarily documenting right-wing paramilitary organizations such as Clan del Golfo and Los Puntilleros. From these reports, we construct *municipio*-level indicators capturing the presence of right-wing armed groups.

The INDEPAZ data primarily document right-wing paramilitary organizations and do not comprehensively capture the activities of left-wing guerrilla groups such as FARC or ELN (*Ejército de Liberación Nacional*; National Liberation Army), both of which have historically played important roles in cocaine trafficking networks. Therefore, to complement the INDEPAZ data, we incorporate an additional dataset on violent events involving armed groups in Colombia. This dataset is constructed using large language models (LLMs) to classify whether specific armed groups were involved in violent incidents described in Colombian news articles, including murders, attacks, kidnappings, and related events (we refer the reader to (Anderson et al. 2025) for additional details on the data construction process). Using these classified news records, we construct annual *municipio*-level indicators for violent activity associated with both left- and right-wing armed groups.

We combine the INDEPAZ data and the violence-event dataset to construct two binary indicators: "left_wing" and "right_paramilitary", which represent the presence of left-wing guerrilla groups and right-wing paramilitary groups, respectively. We aggregate observations from 2013 to 2017, which is a longer aggregation window than the other variables used in the study. Armed groups typically maintain territorial influence over extended periods rather than relocating frequently. The aggregation over a longer period reflects these facts and reduces the sparsity of the data. The variable left_wing is equal to 1 if at least one violent event involving a left-wing armed group occurred in a *municipio*

during the five years, and 0 otherwise. Similarly, right_paramilitary equals 1 if evidence from either dataset indicates the presence or violent activity of a right-wing paramilitary group in the *municipio*. We further introduce an interaction term of left_wing and right_paramilitary, "left_right", to capture *municipios* in which both left- and right-wing armed groups were present, which may indicate regions of heightened strategic importance for cocaine trafficking.

2.2.2.3 Demographic and Geographic Data

We collect publicly available geographic and demographic information for Colombia, including data on rivers, roads, ferry terminals, airports, police stations, military facilities, and population size. These data are obtained from the OpenStreetMap (OSM) project, except for population size from the Colombian National Department of Statistics (DANE 2018).

The OSM's river data classify waterways into four categories: river, stream, canal, and drain. Since major rivers are the most likely to serve as transportation routes for trafficking activities, we retain the "river" category in the analysis. The river data consist of sequential geographic coordinate points connected by line segments. We aggregate these data into a *municipio*-level variable, "river_length", defined as the total Euclidean length of rivers within each *municipio*.

We apply an analogous procedure to the road network data. The road dataset includes motorways, trunk roads, and primary and secondary roads. We retain all roads belonging to these categories and construct the variable "road_length" as the total Euclidean road length within each *municipio*.

Lastly, we use point-location data for ferry terminals, airports, police stations, and military land use; ferry terminals and airports facilitate cocaine transportation, while the presence of police and military may act as a trafficking deterrent. These data are converted into *municipio*-level binary

indicators representing the presence of each facility type within a *municipio*.

2.2.3 Final Dataset

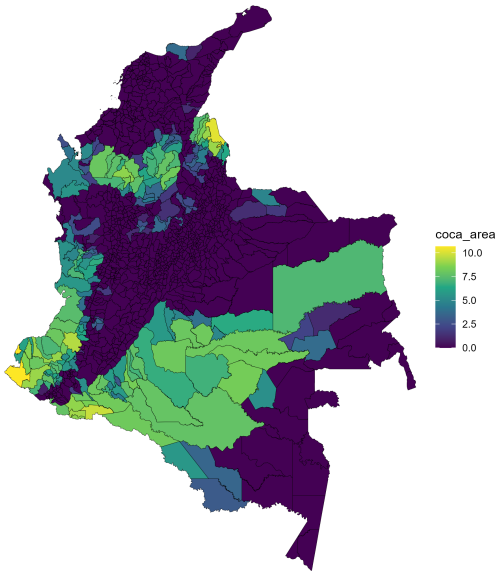
The final regression model includes the following variables: coca cultivation area (coca_area), cocaine hydrochloride seizures (seizures), the predicted probability of cocaine hydrochloride lab presence (lab_prob), the distances of rivers and roads (river_length and road_length), population, binary indicators for right-wing paramilitary presence (right_paramilitary), left-wing guerrilla presence (left_wing), and their interaction term (left_right), and binary indicators of an airport (airport), a ferry terminal (ferry), a police station (police), and military land use (military). These variables are summarized in Table 2.1.

We applied a logarithmic transformation to the coca_area, seizure, and population variables to reduce their skewness. Then, coca_area, seizure, lab_prob, population, river_length, and road_length variables were normalized using their respective means and standard deviations. This normalization sets the variables on comparable scales and improves numerical stability in regression analysis.

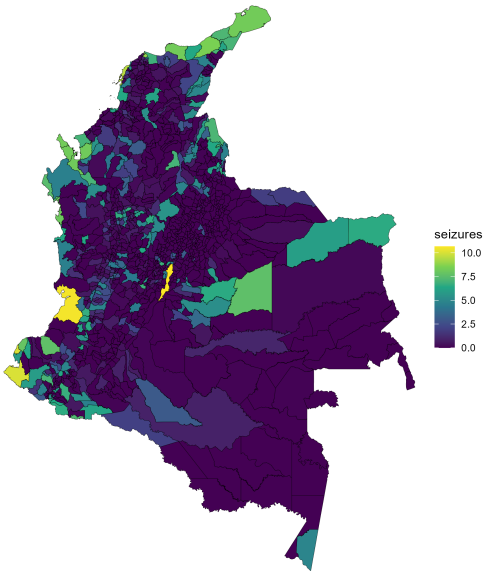
As an illustration, Figure 2.2 presents the spatial variation of lab_reported and the three cocaine-related variables: coca_area, seizures, and lab_prob. Both coca cultivation (coca_area) and the presence of reported cocaine laboratories (lab_reported) exhibit clear spatial clustering in Figures 2.2(a) and 2.2(c), indicating that large cultivation areas and laboratory detections tend to occur in geographically contiguous regions rather than being randomly distributed across Colombia. Coca cultivation areas and cocaine laboratories are concentrated in *municipios* located in southern Colombia, along the Pacific coast, and in parts of north-central Colombia. Seizure activity and lab_prob also display clear spatial patterns. Figure 2.2(b) illustrates that higher seizure levels are concentrated along the western and

Variable Name	Definition	Source
Dependent Variable (y)	A binary indicator of cocaine trafficking destinations in Colombia	Colombian Anti-narcotics Police
Cocaine-Related Variables		
coca_area	Area of coca cultivation (normalized log values)	UNODC (2023)
seizures	Seizure weight of cocaine before the analysis period (normalized log values)	ODC (2022)
lab_prob	Predicted probability of cocaine lab presence	ODC (2022)
Conflict Variables		
left_wing	A binary indicator of left-wing guerrilla groups	Anderson et al. (2025)
right_paramilitary	A binary indicator of right-wing paramilitary groups	Indepaz (2020) Anderson et al. (2025)
left_right	Interaction of left- and right-wing groups	
Geographic and Demographic Variables		
population	Population in a *municipio* (normalized log values)	DANE (2018)
river_length	Total length of rivers in a *municipio* (normalized values)	OpenStreetMap
road_length	Total length of trunk, primary, and secondary roads in a *municipio* (normalized values)	OpenStreetMap
airport	A binary indicator of airport	OpenStreetMap
ferry	A binary indicator of ferry terminal	OpenStreetMap
military	A binary indicator of military land use	OpenStreetMap
police	A binary indicator of police station	OpenStreetMap

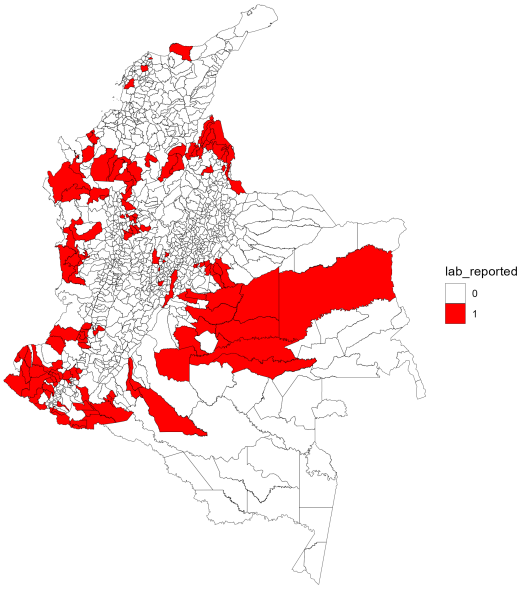
Table 2.1: Variables used in the analysis.



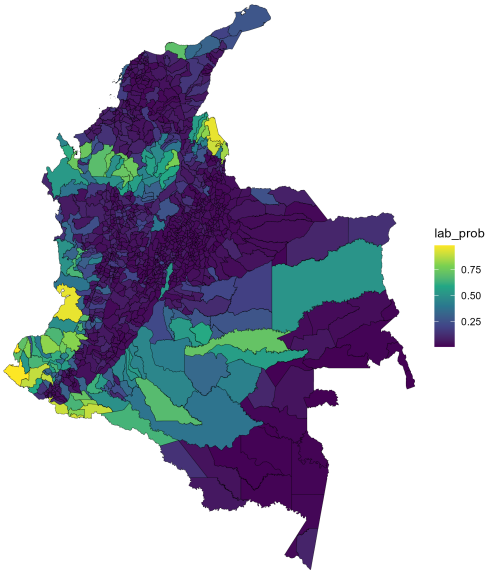
(a) coca_area (log of ha)



(b) seizures (log of kg)



(c) lab_reported



(d) lab_prob

Figure 2.2: Geographical variation of (a) coca_area (log of ha), (b) seizures (log of kg), (c) lab_reported, and (d) lab_prob data.

northern coastlines and in the vicinity of Bogotá D.C. (yellow region in the center). In Figure 2.2(d), lab_prob is distributed with substantial geographic variation, which overlaps most regions with a reported laboratory. Additional maps presented in Appendix A (Figures A.1–A.3) summarize the remaining independent variables, each similarly exhibiting distinct spatial patterns. The observed geographic variation across both the dependent and independent variables further motivates the use of a localized modeling framework to capture spatially varying relationships.

2.3 Methodology

This section presents the methodological framework used to identify region-specific risk factors associated with cocaine hydrochloride destinations in Colombia. The sparsity and imbalance in our data motivate our approach, as such data can lead to overfitting when local models based on relatively small neighborhoods are estimated. To address this challenge, Section 2.3.1 presents the local geographically weighted regression framework with Firth's penalized maximum likelihood estimation. This framework mitigates separation problems that frequently arise when applying a GWR model to sparse data. We also describe the use of binary spatial weights and a bandwidth selection procedure based on predictive performance. Finally, we introduce modeling parameters that ensure the necessary variation across variables.

2.3.1 Local Geographically Weighted Regression with Penalized Maximum Likelihood Estimation

In this study, we chose to use the geographically weighted regression (GWR) approach (Brunsdon et al. 1996, Fotheringham et al. 2002) to capture the potentially geographically diverse relationship

between our dependent variable and the various factors in our data set. Unlike the ordinary regression model, which assume spatially invariant relationships, GWR estimates location-specific coefficients for each focal point, a *municipio* in our case.

In a mathematical form, if there are n *municipios*, the logistic regression equation for location i can be expressed as:

$$P(y_i = 1) = \text{logistic}\left(\beta_{0i} + \sum_{m=1}^{M} \beta_{mi} x_{mi}\right), \tag{2.1}$$

where y_i is a dependent variable, x_{mi} is a value for the mth independent variable of location i, and M is the number of independent variables. Local estimates (β_{mi}) are obtained by only allowing nearby observations to contribute to the estimation process. We refer to our approach as "local" GWR, since, for each focal *municipio*, we only consider nearby locations, *municipios* within a given distance from i to fit the logistic regression model. Let d_{ij} be the Euclidean distance between the centroids of *municipios* i and j, where the distance is computed from their longitude and latitude coordinates.[3] Then the data from municipio j is included when fitting the model for *municipio* i if d_{ij} is less than a bandwidth b. We will refer to the locations for which $d_{ij} \leq b$ as the "*neighbors*" of *municipio* i. We note that the *neighbors* of *municipio* i do not include the observation itself (x_{mi}) to prevent it from influencing its own prediction (see Section 2.3.1.2 for details).

When fitting GWR models, it is common to use continuous weighting schemes such as exponential or Gaussian kernels, such that data closer to each focal point weigh more. We adopt the binary weight approach to highlight local geographic variation in the estimated coefficients more effectively; observations within the local neighborhood contribute equally to model estimation. This choice reflects our assumption that the relationship between cocaine hydrochloride destinations (the dependent variable)

[3] As a reference, one degree of longitude or latitude is approximately 111 km, implying that a bandwidth of b corresponds roughly to $111 \times b$ km.

	Cocaine Destination: Yes	Cocaine Destination: No
Airport: Yes	9	0
Airport: No	27	56

Table 2.2: An example of quasi-complete separation in the dependent variable (Cocaine Destination), with the binary indicator for airport's presence in a *municipio* (Airport).

and the independent variables varies spatially, and only local data near a *municipio* are relevant.

For a *municipio* i, the best bandwidth, b_i^*, can be dynamically determined based on any goodness of fit or performance measure. We use the Area Under the Curve (AUC) as our measure and vary the value of b_i^* from 0.5 to 4.0 in steps of 0.1, to determine b_i^*, the best bandwidth for each *municipio*.

2.3.1.1 Quasi-Complete Separation in Observations

Complete or quasi-complete separation refers to one or more of the independent variables perfectly or nearly perfectly predicting a binary outcome, which in turn leads to non-convergence or infinite coefficient estimates (Heinze and Schemper 2002). As an example, consider Table 2.2 in which the variable airport only takes the value 1 for *municipios* with a positive outcome, and is either 1 or 0 for all *municipios* with a negative outcome. This is an example of quasi-complete separation, which undermines the stability and reliability of parameter estimates and can lead to inflated regression coefficients.

The separation issue becomes especially prominent when the bandwidth is small or when fitting models for *municipios* in southeast Colombia, where *municipios* tend to be geographically larger, resulting in a fewer number of *neighbors* within a given bandwidth. We implement Firth's penalized maximum likelihood (PML; Firth 1993) to address this separation issue in our GWR models, as suggested by Heinze and Schemper (2002). Details about Firth's method are described in Appendix B. We will refer to this approach as "local GWR PML".

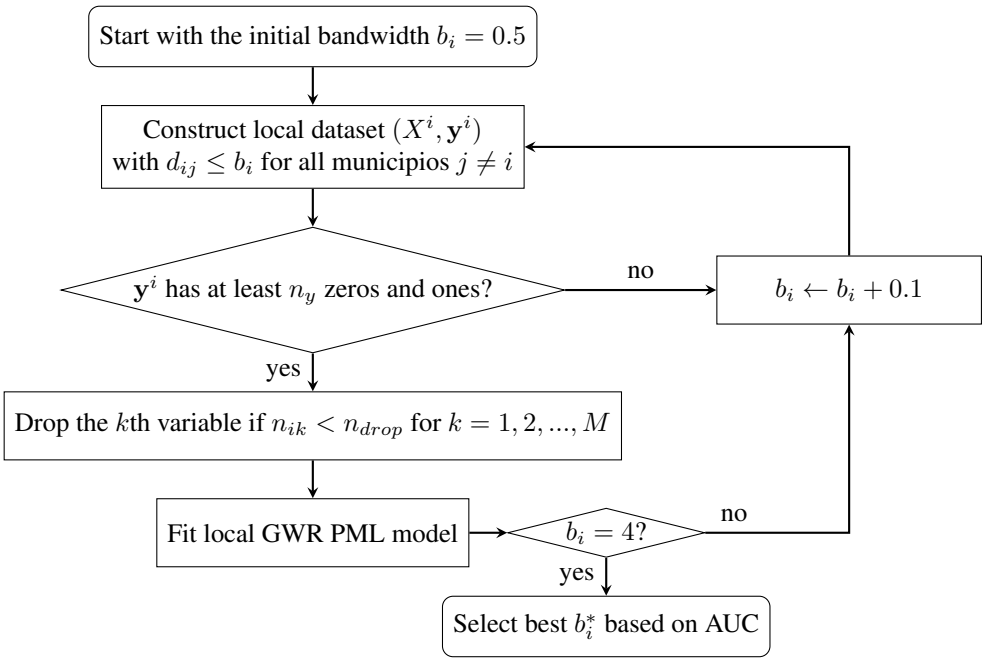


Figure 2.3: Flowchart for local GWR PML model fitting with the variable-drop procedure for *municipio* i.

d_{ij} : the Euclidean distance between municipios i and j,
n_y : the minimum number of observations required in each response class,
M : the number of independent variables,
n_{ik} : the number of distinct values for a continuous variable or the minimum count of occurrences for each class of a binary variable, and
n_{drop} : the threshold used to exclude an independent variable.

2.3.1.2 Implementation of Local GWR PML

At a high level and as summarized in Figure 2.3, we determine the best bandwidth b_i^* by iterating from 0.5 to 4 in steps of 0.1 for each *municipio*, each time fitting a logistic regression model. To avoid basing the optimal bandwidth b_i^* on overfitted models due to the separation issue, we take steps to ensure sufficient data density before fitting any local model. We achieve this goal by requiring enough observations from both classes of the dependent variable and dropping independent variables when the variation in their values is below a threshold.

We denote *neighbors*' independent variables of *municipio* i by X^i, an $N^i \times M$ matrix whose

rows consist of vectors of *neighbors*' observations, where N^i is the number of *neighbors* of i. Similarly, an $N^i \times 1$ vector $\mathbf{y}^i$ denotes the outcomes of *municipio* i's *neighbors*. For each location i and a given bandwidth b_i, the dataset $(X^i, \mathbf{y}^i)$ consists of all *neighbors* whose $d_{ij} \leq b_i$.

We introduce two threshold parameters, n_y and n_{drop}, to ensure reliable model estimation. First, n_y determines if there is a sufficient number of observations for each class in the outcome variable. For a given bandwidth b_i, if either class in $\mathbf{y}^i$ contains fewer than n_y observations, we do not fit the local GWR PML model and increase b_i by 0.1. Second, n_{drop} is used as a threshold for independent variables. If, for a given bandwidth b_i, a numerical variable has fewer than n_{drop} unique values, then the variable is not included in the model. If, on the other hand, the independent variable is binary, then the variable is excluded if either category (0 or 1) occurs fewer than n_{drop} times. Let n_{ik} be the number of distinct values for a continuous variable or the minimum count of occurrences for each class of a binary variable. Variable k is included in the local model only if $n_{ik} \geq n_{drop}$. Once the process is repeated for the full range of bandwidths, we select the best bandwidth corresponding to the highest AUC.

We set $n_y = 8$ and $n_{drop} = 10$ to generate the results of our local GWR PML models. We note that the value of these thresholds is a modeling choice; we report on sensitivity analysis investigating their impact in Section 2.4.4. For each *municipio*, we implemented the local GWR PML model in R programming language version 4.4.2 using the logistf package.

2.3.1.3 Model Evaluation

Our goal of applying the local GWR PML model to our Colombian cocaine trafficking data is twofold. First, we seek to reliably estimate and explain the effects of independent variables on the probability of a *municipio* being a cocaine destination. Second, we aim to identify potential trafficking

nodes not reported in the anecdotal evidence using the model in a "predictive" manner. It is, therefore, important not only to avoid overfitting since it may lead to misleading interpretations or inaccurate predictions but also to evaluate the models' predictive performance. To this end, we select the optimal bandwidths based on the models' performance evaluated using neighbors only. Then, when reporting the predictive performance of the finalized model, we evaluate it on the focal location itself, which was not used in any modeling decisions (e.g., the model fitting and performance estimation).

2.4 Results and Discussion

In this section, we present key findings from the analysis. First, we summarize the impact of different features across Colombia in Section 2.4.1. Then, to illustrate the spatial heterogeneity in the estimated coefficients, we examine the local estimation results for two *municipios*-San Antero (Córdoba) and Convención (Norte de Santander)-as a case study in Section 2.4.2. We discuss two *municipios* identified as potential cocaine destinations in Section 2.4.3. Lastly, Section 2.4.4 presents the results of the sensitivity analysis for the two modeling parameters, n_y and n_{drop}.

2.4.1 Spatial Variation of Regression Coefficients

Figure 2.4 displays the distribution of the estimated local coefficients for all independent variables using violin box plots. To improve visualization of the central distributions, coefficient estimates with absolute values greater than 20 are excluded from the figure. The variable lab_prob has 2.2% of its regression coefficient values outside the displayed range (24 out of the 1,101 *municipios* that include lab_prob variables in their local models), and the variable road_length has one coefficient exceeding this range. For other variables, the figure includes all values.

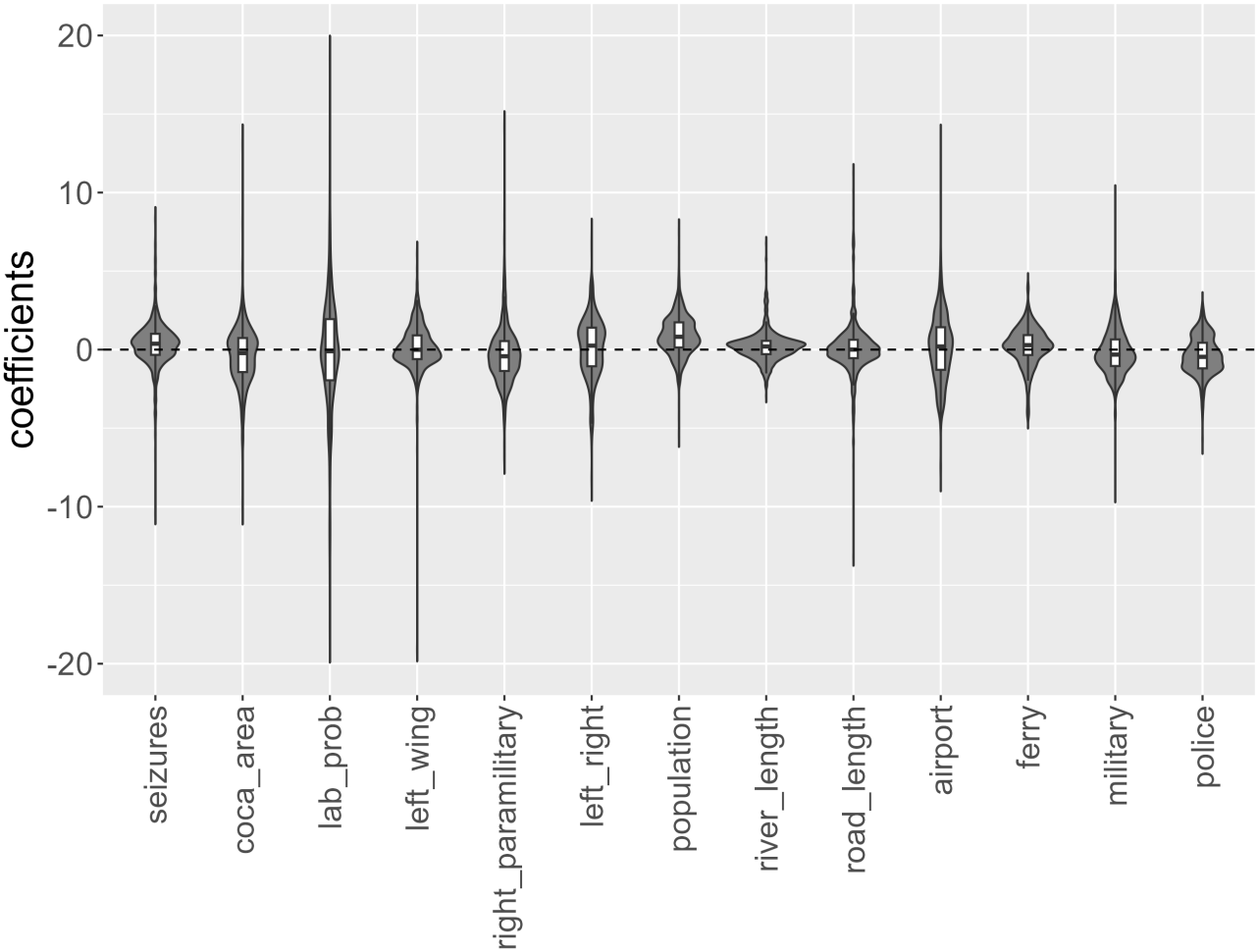


Figure 2.4: Violin box plot of estimated regression coefficients using the local GWR PML model (absolute values greater than 20 are excluded).

Table 2.3: Summary of regression coefficients obtained with the local GWR PML and the global model.
n_{sig}: the number of *municipios* with statistical significance ($p \leq 0.1$) for each variable.

Variable	Global	Local GWR PML Coefficients						
	Coefficients	Min	Q1	Median	Mean	Q3	Max	n_{sig}
Intercept	-0.221	-23.731	-2.595	-1.120	-0.494	0.916	36.284	446
coca_area	-0.839 **	-9.851	-1.432	-0.219	-0.376	0.737	13.055	61
seizures	-0.015	-10.448	-0.334	0.374	0.347	1.008	8.405	235
lab_prob	0.950 **	-64.176	-1.954	-0.058	0.313	2.022	82.535	181
left_wing	-0.252	-19.090	-0.605	-0.008	0.156	0.907	6.113	48
right_paramilitary	-0.518 **	-6.901	-1.361	-0.423	-0.340	0.543	14.172	140
left_right	0.384	-8.295	-1.060	0.263	0.130	1.399	20.078	36
population	0.507 ***	-5.405	0.135	0.815	0.896	1.735	7.505	360
river_length	0.282 ***	-2.941	-0.289	0.200	0.197	0.557	6.758	243
road_length	-0.045	-13.183	-0.543	0.001	0.093	0.628	30.225	279
airport	-0.029	-7.568	-1.287	0.199	0.115	1.423	12.867	307
ferry	0.352 *	-4.192	-0.341	0.283	0.211	0.922	4.037	23
military	0.032	-8.841	-1.044	-0.318	-0.158	0.649	9.579	137
police	-0.525 ***	-5.835	-1.194	-0.461	-0.386	0.435	2.852	209
AUC	0.692	0.825						
Significance	*** p≤0.001	** p≤0.05	* p≤0.1					

A notable feature of Figure 2.4 is that the coefficient distributions for all variables are distributed above and below zero. This pattern indicates again substantial spatial heterogeneity in the relationships between the independent variables and the likelihood of a *municipio* being a cocaine hydrochloride destination. Positive coefficients are most prevalent for seizures, population, river_length, and ferry, whereas negative coefficients are observed predominantly for right_paramilitary, military, and police. Lastly, lab_prob exhibits notably wider coefficient distributions than the other covariates. This variation indicates that the relationship between predicted laboratory presence and trafficking destinations differs substantially across Colombia, suggesting that cocaine production laboratories may play different roles within local trafficking networks.

To summarize the coefficient distributions shown in Figure 2.4, Table 2.3 lists descriptive statistics

of the estimated coefficients alongside the corresponding global regression coefficients for comparison, estimated with a single model using data from all *municipios*. For each variable, Table 2.3 reports the global coefficient followed by the local coefficients' minimum, first quartile (Q1), median, mean, third quartile (Q3), and maximum values across *municipios*.

Consistent with the objectives of the local modeling framework, the coefficients estimated for the local GWR PML models differ substantially from the global estimates, including some *municipios*' estimates having the opposite coefficient signs. These wide ranges of estimates demonstrate the ability of the GWR approach to capture spatial heterogeneity in the relationships between the independent variables and cocaine trafficking destination.

We compute the AUC values for both the global and local GWR PML models by predicting the dependent variable (y_i) for each *municipio* using the estimated coefficients and covariate values at location i. For the local GWR PML model, these predictions are generated from models estimated without the focal location. The local GWR PML model achieves an AUC that is 0.133 higher than that of the global model (0.825 versus 0.692), indicating a better predictive capability than the global model.

The geographically heterogeneous effects identified by the local GWR PML model can be further studied through the statistical significance of the estimated coefficients (variables with a p-value less than 0.1). The n_{sig} value in Table 2.3 counts the number of times a regression coefficient was statistically significant. A considerable number of coefficients have consistently significant effects on the outcome. The n_{sig} value also illustrates how geographically weighted regression can reveal relationships that are hidden in a global model. As shown in Table 2.3, the global regression identifies coca_area, lab_prob, right_paramilitary, population, river_length, ferry, and police as statistically significant independent variables. In contrast, the three variables with the largest n_{sig} in the local GWR PML models are population (360), airport (307), and road_length (279), only one of which is significant in the global

model. This difference suggests that the effects of these variables vary considerably across space, with positive, negative, or negligible associations depending on local conditions. By allowing coefficients to vary geographically, the GWR framework is able to uncover localized relationships with cocaine trafficking destinations that may be masked when a single coefficient is estimated for the entire study area.

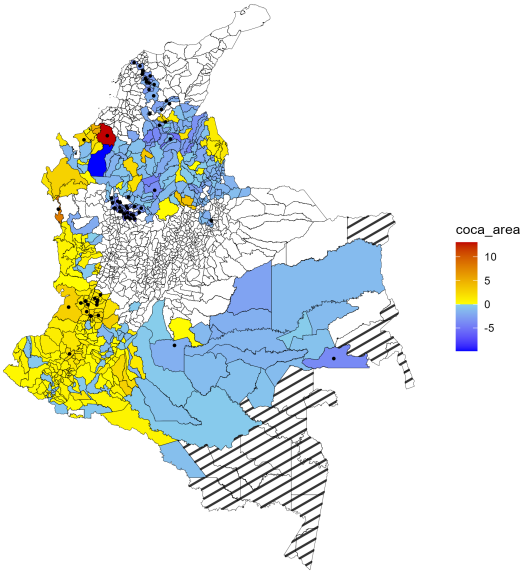
2.4.1.1 Cocaine-Related Variables

Figure 2.5 presents the estimated coefficients for three cocaine-related variables —coca_area, seizures, and lab_prob— obtained from the local GWR PML model. To emphasize statistically meaningful spatial patterns, black points in *municipio* centroids indicate that the estimated coefficient for a given variable is statistically significant at the 10% level.

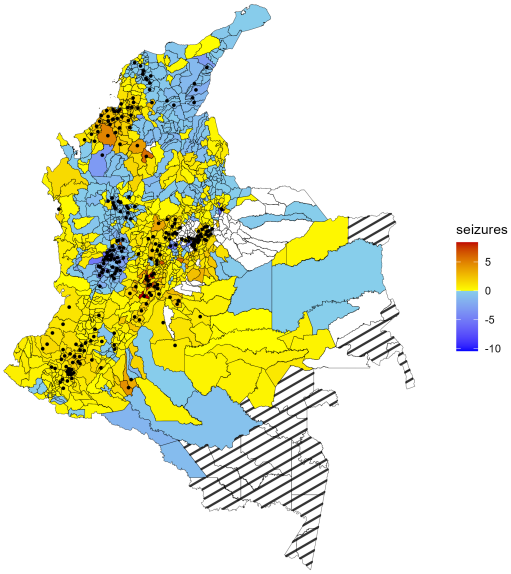
Note that coefficient estimates are unavailable for *municipios* in southeastern Colombia. Although larger bandwidths could increase the number of neighboring *municipios* available for estimation, it would require incorporating data from geographically distant *municipios*, which is inconsistent with the underlying assumption of GWR that local relationships are best characterized by nearby locations. Consequently, we restrict estimation to bandwidths that preserve the local nature of the model, resulting in 19 *municipios* being excluded from the analysis.

The regression coefficients summarized in Figure 2.5 reveal substantial regional heterogeneity, indicating that the relationships between these variables and the probability of a *municipio* serving as a cocaine hydrochloride destination vary considerably across Colombia.

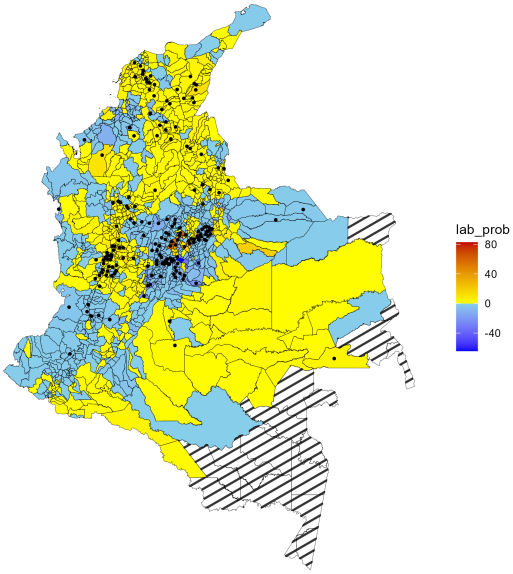
Positive coefficients for coca_area along the western coastline of Colombia, in Figure 2.5(a), suggest that coca cultivation in these regions is associated with a higher probability of a *municipio* being



(a) coca_area



(b) seizures



(c) lab_prob

Figure 2.5: Estimated regression coefficients of (a) coca_area, (b) seizures, and (c) lab_prob using the local GWR PML model.

Notes: Positive associations are represented by colors ranging from yellow to red according to their magnitude, whereas negative coefficients are shown in shades of light to dark blue. *Municipios* with coefficients equal to zero are shaded in gray, while white regions indicate locations where the corresponding variable was excluded from the local GWR model due to insufficient variation in its values among *neighbors*. Striped regions are excluded from the analysis.

a cocaine hydrochloride trafficking destination (e.g., Cauca, Chocó, and Valle del Cauca). This pattern may reflect the presence of more established and spatially integrated cultivation-to-manufacturing supply chains in these areas, potentially indicating greater organizational coordination in cocaine production and trafficking activities relative to other parts of the country (Kenney 2007). In contrast, coca cultivation exhibits weak, negligible, or in some cases negative associations in several inland regions (e.g., Antioquia, Bolívar, and the southeastern areas), highlighting the substantial spatial heterogeneity in the role of coca cultivation within Colombia's cocaine economy. These variations may reflect a divergence in illicit supply chain structures managed by distinct transnational criminal organizations. While some networks exhibit high levels of vertical integration, others operate within more fragmented, market-driven frameworks, with structures heavily influenced by their strategic orientation toward specific transit corridors, such as the Pacific maritime, Caribbean littoral, or contiguous land borders.

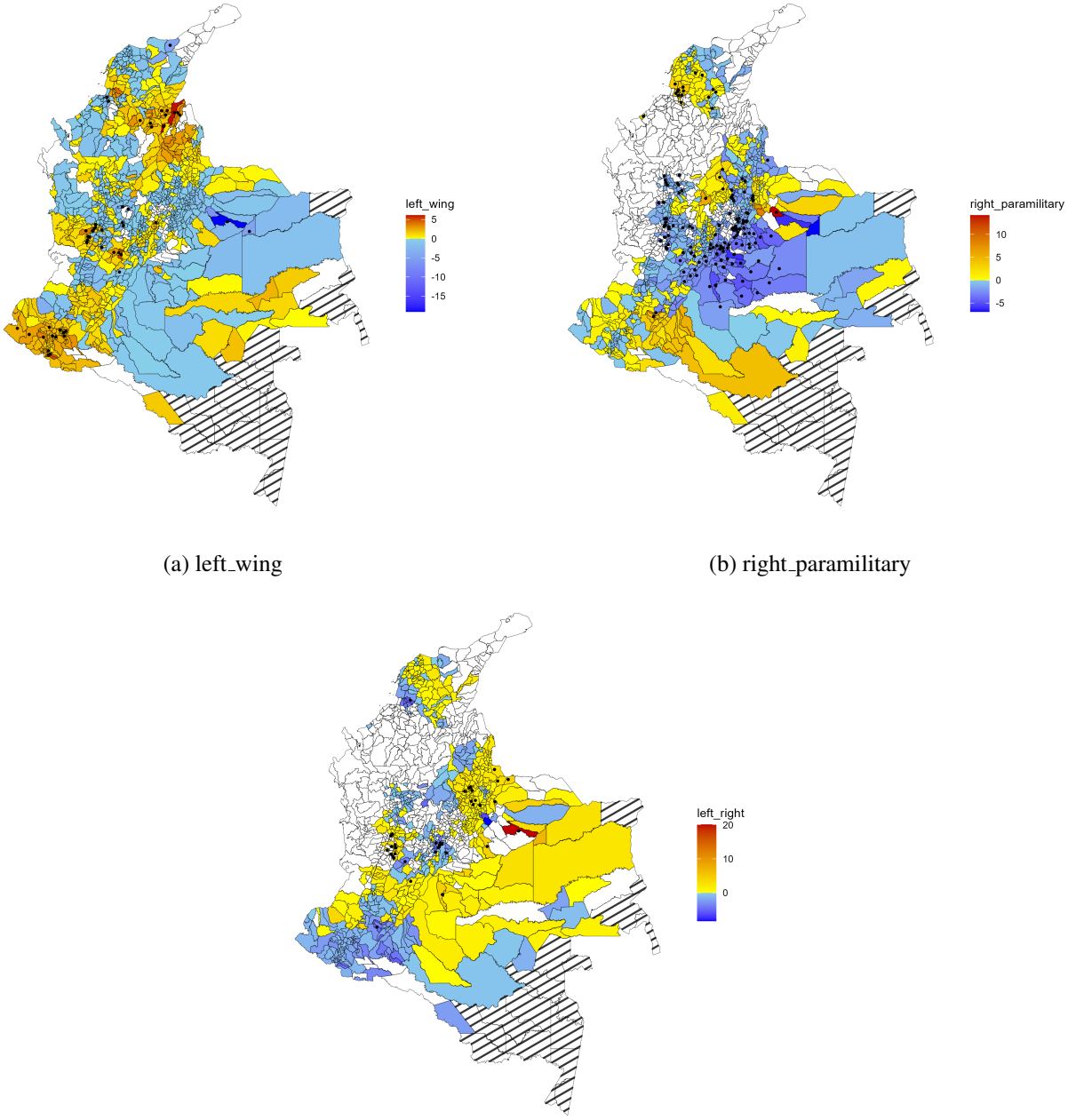
The estimated coefficients for seizures in Figure 2.5(b) exhibit more positive patterns across much of Colombia, with positive associations observed in nearly all west-coast *municipios*. Since the seizure variable represents the quantity of cocaine confiscated leading up to the analysis period, positive coefficients suggest that *municipios* with prior seizure records are, on average, more likely to continue functioning as active cocaine trafficking destinations. At the same time, several regions have negative seizure coefficients (e.g., Atlántico and Risaralda). These areas may reflect the so-called "balloon effect" of law enforcement interventions, in which intensified enforcement pressure in one location displaces trafficking activity toward alternative routes or neighboring regions (Rouse and Arce 2006). An alternative explanation is that the peace process influenced cocaine trafficking away from these areas. The observed spatial heterogeneity suggests that seizures may capture stable trafficking corridors in some parts of the country while indicating deterrence or displacement effects in others. These geographically varying relationships provide potentially important implications for the design of localized enforcement

and interdiction strategies. Specifically, positive seizure coefficients are concentrated along Colombia's Caribbean coast of Córdoba and Sucre, the major Andean corridor connecting Bogotá D.C. to northern export points, and portions of the Pacific coast, regions corresponding to established high-volume trafficking corridors.

As shown in Figure 2.5(c), the estimated coefficients of lab_prob are moderate across most regions, except in central Colombia (Cundinamarca). Moderately positive coefficients are concentrated primarily in northern and southeastern Colombia and tend to coincide with areas exhibiting high predicted probabilities of cocaine processing lab presence in Figure 2.2(d). Conversely, *municipios* with negative coefficients generally have lower predicted probabilities of laboratory presence. Several regional exceptions to this rule are observed. Along parts of the Pacific coast, particularly in Cauca and Nariño, higher lab_prob values are associated with a lower likelihood of being a cocaine destination. Further, parts of the Caribbean region in northern Colombia, including Atlántico, northern Bolívar, and Cesar, exhibit positive coefficient values despite relatively low lab_prob values. These contrasting patterns suggest that the relationship between cocaine processing laboratories and trafficking destinations varies across Colombia.

2.4.1.2 Conflict Variables

We next examine how the presence of two major armed groups, left-wing guerrillas (e.g., FARC and ELN) and right-wing paramilitary groups (e.g., Clan del Golfo and Los Puntilleros), is associated with cocaine trafficking destinations. The local coefficient estimates for left_wing and right_paramilitary obtained from the local GWR PML model are mapped in Figures 2.6(a) and 2.6(b), respectively. Overall, positive associations between cocaine trafficking destinations and the presence of left-wing groups are



(a) left_wing

(b) right_paramilitary

(c) left_right

Figure 2.6: Estimated regression coefficients of (a) left_wing, (b) right_paramilitary, and (c) left_right using the local GWR PML model.

Notes: Positive associations are represented by colors ranging from yellow to red according to their magnitude, whereas negative coefficients are shown in shades of light to dark blue. *Municipios* with coefficients equal to zero are shaded in gray, while white regions indicate locations where the corresponding variable was excluded from the local GWR model due to insufficient variation in its values among *neighbors*. Striped regions are excluded from the analysis.

concentrated mainly in the southern Pacific coast of Nariño *departamento* and the Northeast of Colombia (e.g., Cesar, Norte de Santander) bordering Venezuela. In central Colombia (e.g., Cundinamarca and Tolima), several *municipios* exhibit positive coefficients for left_wing. In contrast, positive coefficients for right_paramilitary are relatively uncommon across the central region and are generally small in magnitude. This contrast is consistent with prior research documenting strong links between FARC presence and the cocaine business, as well as important differences between left-wing guerrilla and right-wing paramilitary organizations (González 2022, Gutiérrez Sanín 2008, Mateus-Rugeles and Arias 2020, Norman 2018, Sierra 2016). However, some *municipios* along the northern coastline of Colombia (e.g., Magdalena and northern Bolívar) exhibit positive associations with right_paramilitary and negative or weaker associations with left_wing.

These findings suggest that the relationships between armed groups and cocaine trafficking destinations vary considerably across Colombia. Such spatial heterogeneity is consistent with the decentralized nature of Colombia's illicit drug economy, in which armed organizations may differ in their territorial influence, operational strategies, available resources, and involvement in trafficking activities across regions (Grajales-Marín et al. 2025, Kenney 2007). The contrasting coefficient signs observed in central Colombia and along the northern coast further indicate that the relative importance of different armed actors may vary geographically. Consequently, analyses based solely on global models may obscure important local dynamics linking armed conflict and illicit trafficking.

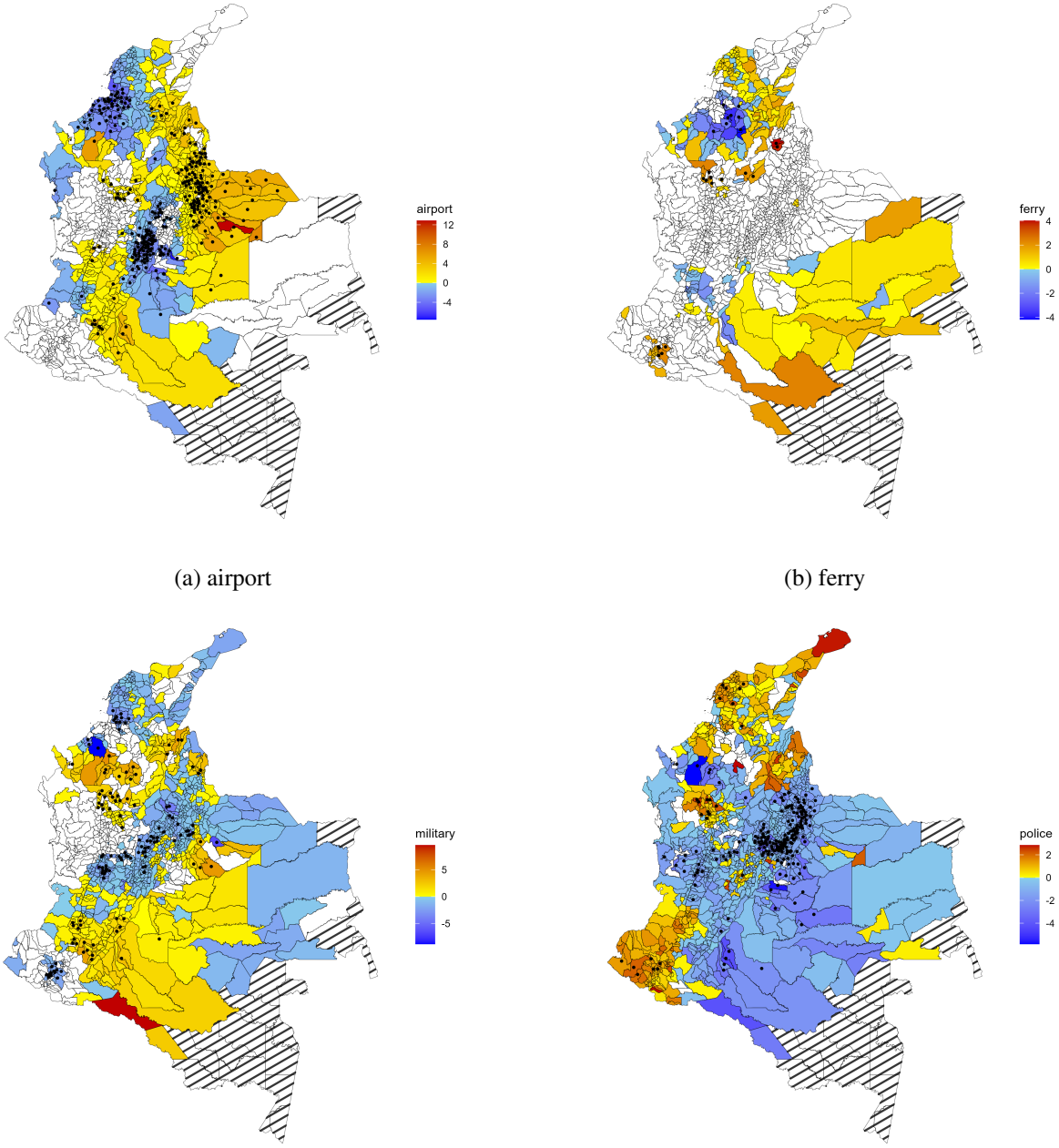
To further investigate the joint presence of armed groups, we added an interaction term between left_wing and right_paramilitary, with the resulting local coefficients shown in Figure 2.6(c). Interestingly, several *municipios* in the central and northern regions (e.g., Boyacá and Santander) that exhibit negative coefficients for both groups individually display positive coefficients with their joint presence. Conversely, in parts of southern Colombia (e.g., Caquetá), where each group individually shows

a positive association with trafficking destinations, the interaction term is negative. These results suggest that the combined presence of the two groups may be associated with trafficking destinations in ways that differ from their individual effects, and that such relationships are themselves spatially heterogeneous. The observed patterns may reflect regional differences in competition, coexistence, or other forms of interaction between different armed groups.

2.4.1.3 Infrastructures, Rivers, and Population

Infrastructure-related variables (roads, airports, ferry terminals, military land use, and police stations), as well as rivers and population, again exhibit substantial spatial heterogeneity in their relationships with cocaine trafficking destinations, as summarized in Figures 2.7 and 2.8. The spatially varying coefficients for airports and ferry terminals (Figures 2.7(a) and 2.7(b)) suggest that the likelihood of being a cocaine trafficking destination varies considerably across Colombia.

Positive and statistically significant coefficients suggest regions where air or ferry transportation may play a more important role in cocaine trafficking than other regions. These localized patterns can help guide future investigations into the mechanisms through which different transportation infrastructures facilitate trafficking activity. For example, Bogotá D.C., Cundinamarca, and Sucre with weak or negative airport coefficients may indicate that other modes of transport are more influential in the area. One may further examine such areas to identify factors underlying the reduced influence of airports on trafficking outcomes. Potential explanations include differences in airport type, traffic volume, ownership structure, and security or inspection capacity. Such analyses could provide valuable insights into the factors driving regional variation in the role of airports and ferry ports within cocaine trafficking networks.



(a) airport

(b) ferry

(c) military

(d) police

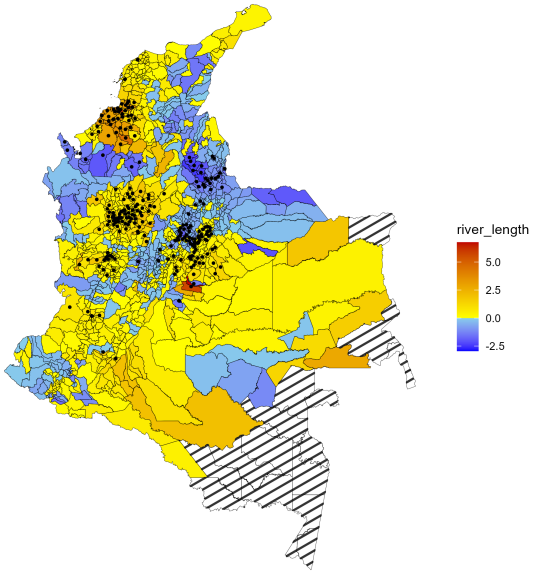
Figure 2.7: Estimated regression coefficients of (a) airport, (b) ferry, (c) military, and (d) police using the local GWR PML model.

Notes: Positive associations are represented by colors ranging from yellow to red according to their magnitude, whereas negative coefficients are shown in shades of light to dark blue. *Municipios* with coefficients equal to zero are shaded in gray, while white regions indicate locations where the corresponding variable was excluded from the local GWR model due to insufficient variation in its values among *neighbors*. Striped regions are excluded from the analysis.

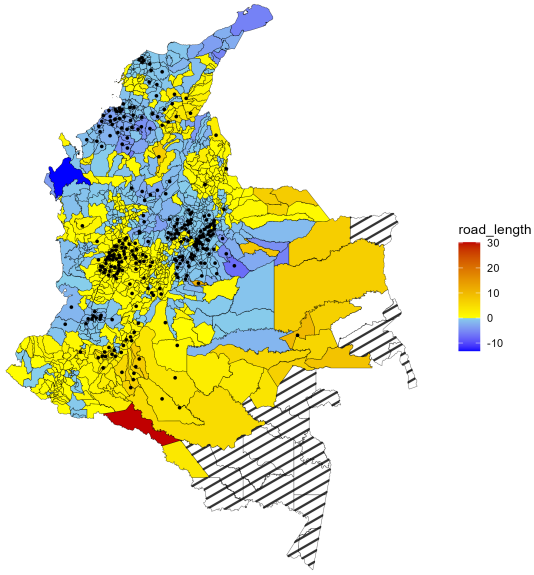
The coefficients for military land use and police stations, summarized in Figures 2.7(c) and 2.7(d), are mostly negative across much of the country, consistent with the expectation that the presence of law enforcement acts as a deterrent for trafficking activity (Du et al. 2026, Mejia and Restrepo 2016). Nevertheless, several areas, such as the west coast and northern Colombia (e.g., Atlántico, La Guajira, and Nariño), exhibit positive coefficients for police. Such patterns may arise for multiple reasons. One possibility is that regions experiencing elevated trafficking activity receive greater allocations of military and police resources, resulting in a positive association that reflects enforcement responses rather than a direct deterrent effect. Alternatively, these coefficients may indicate regional differences in enforcement effectiveness or the adaptive behavior of trafficking organizations. Consequently, caution is warranted when interpreting these relationships causally or when using them for predictive purposes in regions where positive coefficients are observed.

The estimated coefficients for river and road lengths (Figures 2.8(a) and 2.8(b)) also display distinct spatial variation. In particular, statistically significant positive coefficients indicate regions where rivers or road networks may serve as trafficking routes to cocaine destinations, suggesting that they may facilitate the transportation of cocaine in those areas. In contrast, negative coefficients observed elsewhere, for example in northwestern Colombia, imply that the importance of rivers and roads is not uniform across the country. Such differences may reflect regional variation in accessibility, transportation efficiency, geographic constraints, or the relative risk of interdiction along specific routes.

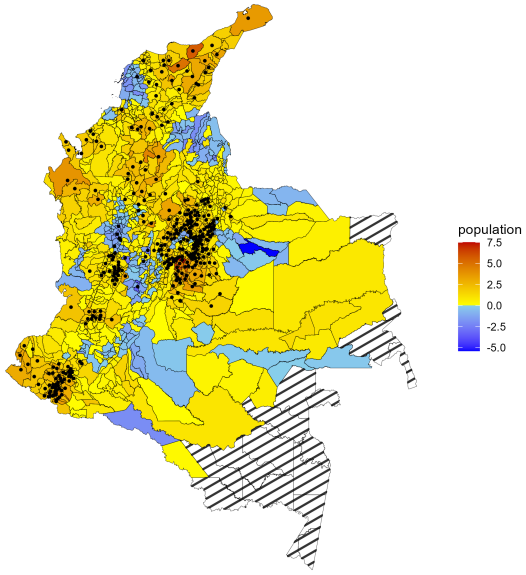
Finally, Figure 2.8(c) highlights that population is positively associated with cocaine trafficking destinations in most regions of Colombia. This pattern is consistent with the tendency for trafficking destinations to be located near major population centers, which may function as consumer markets, logistical hubs, or gateways to export routes. However, several *municipios* exhibit negative population coefficients. These locations may represent less conspicuous trafficking routes, intermediate transshipment



(a) river_length



(b) road_length



(c) population

Figure 2.8: Estimated regression coefficients of (a) river_length, (b) road_length, and (c) population using the local GWR PML model.

Notes: Positive associations are represented by colors ranging from yellow to red according to their magnitude, whereas negative coefficients are shown in shades of light to dark blue. *Municipios* with coefficients equal to zero are shaded in gray, while white regions indicate locations where the corresponding variable was excluded from the local GWR model due to insufficient variation in its values among *neighbors*. Striped regions are excluded from the analysis.

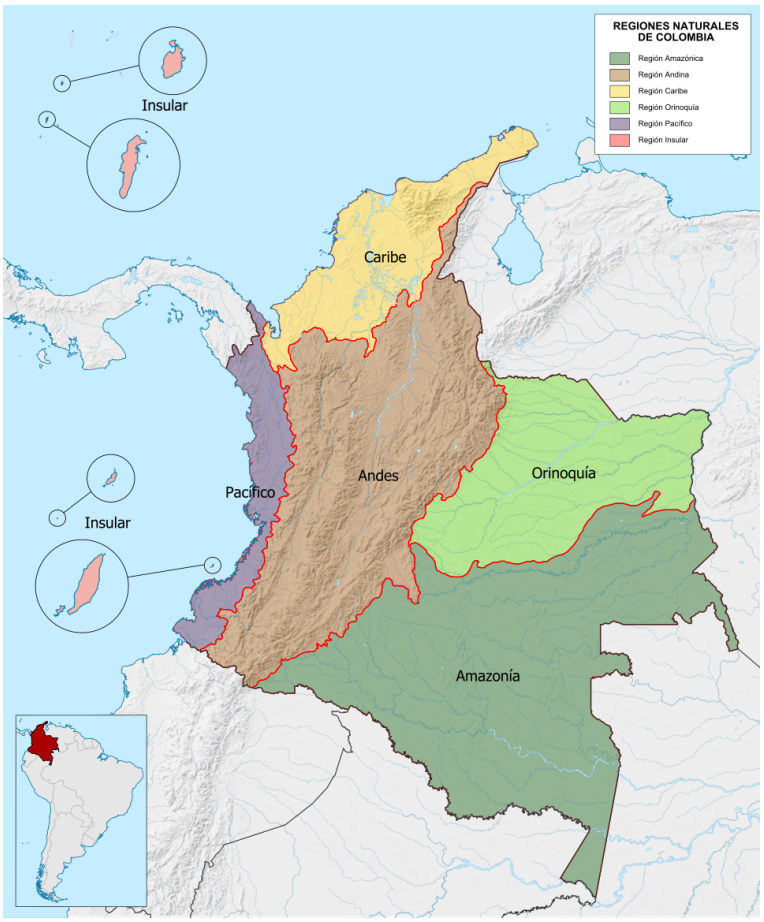


Figure 2.9: Natural regions of Colombia (Milenioscuro 2014).

points, or smuggling routes that avoid major population centers in order to reduce interdiction risk.

To summarize the geographically weighted regression results presented in Figures 2.5–2.8, we organize the findings according to Colombia's natural and geographic characteristics. Mainland Colombia is commonly divided into five natural regions (Rivera 2013): Amazon (Amazonía), Andean (Andes), Caribbean (Caribe), Orinoquía, and Pacific (Pacífico), as depicted in Figure 2.9. Population and river_length show relatively consistent positive associations with trafficking destinations across much of the country, whereas the effects of other factors vary considerably by region.

In the Pacific region, particularly Cauca and Valle del Cauca, coca cultivation is a dominant positive factor. Within the Andean region, river accessibility is positively associated with trafficking

Variables	San Antero		Convención	
	Value	Coefficients	Value	Coefficients
Intercept		4.440		-6.450 ***
coca_area	-0.416		2.960	-2.150
seizures	2.660	2.090 *	0.020	-0.857
lab_prob	0.281	-4.86	2.530	2.880
left_wing	1	-1.16	1	3.420 **
right_paramilitary	1		1	
left_right	1		1	
population	0.707	0.943	0.305	-0.966
river_length	-0.420	3.530 **	0.435	-0.222
road_length	-0.131	-2.240 **	-0.578	-0.702
airport	0	-3.520 **	1	2.230
ferry	1		0	0.933
military	1		1	2.430 *
police	1	0.800	0	-0.232
Significance	*** p≤0.001	** p≤0.05	* p≤0.1	

Table 2.4: Local GWR PML model results of San Antero, Córdoba, and Convención, Norte de Santander.

destinations in northern areas such as Antioquia and Boyacá, while road accessibility is more important in central areas such as Caldas, Risaralda, and Tolima. In the Caribbean region, police station presence is strongly associated with destinations in Atlántico, whereas river length is more influential in Córdoba. In the Amazon region, both river and road accessibility are positively associated with trafficking destinations, while airport presence is a consistent positive factor in the Orinoquía region.

2.4.2 Case Study: San Antero, Córdoba, and Convención, Norte de Santander

To highlight the geographic heterogeneity captured by the local GWR PML model, we compare the estimated coefficients for the two *municipios* with high predicted probabilities of being cocaine destinations: San Antero, Córdoba, located on Colombia's northern Caribbean coast, and Convención, Norte de Santander, located in northeastern Colombia near the Venezuelan border. Both *municipios* are

observed cocaine destinations ($y_i = 1$) and their predicted probabilities are 1 and 0.931, respectively.

Although both locations are assigned similarly high predicted probabilities by the model, the factors driving these predictions differ substantially. As summarized in Table 2.4, eight variables are included in both local models: seizures, lab_prob, left_wing, population, river_length, road_length, airport, and police. In contrast, coca_area, ferry, and military are included in the Convención's model.

In San Antero, seizures constitute the primary contributor to the high predicted probability. The combination of a large value and a statistically significant positive coefficient of seizures produces the largest positive contribution among the selected variables. River and road_length also exhibit significant but opposing effects, with waterways positively associated with destination status and roads negatively associated. However, because San Antero, Córdoba has negative values for both variables (recall that the data are normalized), limited road infrastructure contributes positively to the prediction, while the lack of rivers to use contributes negatively. The airport coefficient is significantly negative, but because no airport is present in the *municipio*, it does not affect the predicted probability. Overall, the results suggest that trafficking activity in San Antero is primarily associated with prior seizure activity. These seizures fail to act as an effective deterrence because this *municipio* serves as a critical export hub within the drug trafficking network, dominated by the Clan del Golfo.

In contrast, the presence of left_wing guerrillas and military land use emerges as the dominant factors in Convención, Norte de Santander. Both variables have statistically significant positive coefficients and generate positive contributions to the predicted probability. Unlike San Antero, seizures, river, and road_length are not statistically significant, indicating a different local trafficking context. Despite a strongly negative intercept, which suggests a relatively low baseline probability of trafficking activity in surrounding *municipios*, the presence of left_wing armed groups and military land use substantially increases the likelihood that Convención is a cocaine destination. These results suggest

that armed groups operating in the area may be closely connected to cocaine trafficking activity, while the presence of military land use may reflect the concentration of regular army in response to their illicit activity.

Taken together, these two cases demonstrate that *municipios* with similarly high probabilities of being cocaine destinations may be characterized by fundamentally different local conditions. While trafficking risk in San Antero appears to be historically associated with interdiction activity, the prediction in Convención is driven primarily by the existence of left_wing groups and the military. These contrasting results illustrate the value of geographically weighted modeling for identifying region-specific drivers that would be obscured by a single global model in Table 2.3. Empirically, these findings highlight the distinct roles that each *municipio* plays within the broader trafficking network, which carries direct implications for antinarcotic operations. San Antero functions primarily as a transit hub to the Caribbean Sea via the Gulf of Morrosquillo, whereas Convención operates as a production and logistics hub situated within the mountainous Catatumbo region along the Venezuelan border. Because both locations present unique geographic particularities and operational challenges, they demand a strategic, highly localized approach rather than a general enforcement model. Consequently, for a production hub like Convención, law enforcement must concentrate on monitoring armed groups' activities; conversely, in a coastal transit node like San Antero, the highest-impact strategies rely on maritime interventions and targeted raids on local storage facilities.

Variables	Medio San Juan		Cumaral	
	Value	Coefficients	Value	Coefficients
Intercept		-1.480		0.030
coca_area	2.250	-0.657	-0.416	
seizures	-0.480	-2.010	-0.450	
lab_prob	1.690	1.940	-0.410	-20.700 **
left_wing	0	1.930	0	
right_paramilitary	1		0	
left_right	0		0	
population	-0.260	-0.850	0.445	2.440
river_length	1.750	2.490 **	0.192	-1.310
road_length	-0.685	-0.330	-0.433	11.200 **
airport	1		1	
ferry	0		0	
military	0		0	
police	0	-2.70 **	1	
Significance	*** p≤0.001	** p≤0.05	* p≤0.1	

Table 2.5: Local GWR PML model results of Medio San Juan, Chocó, and Cumaral, Meta.

2.4.3 Case Study: Potential Cocaine Trafficking Destinations Based on the Local GWR PML Model

To further illustrate the practical value of our local GWR PML framework, we examine two *municipios* with the high predicted probabilities among locations that are not reported as cocaine destinations in the anecdotal data: Medio San Juan, Chocó, located along Colombia's Pacific coast, and Cumaral, Meta, located in the central region. Their predicted probabilities are 0.998 and 0.989, respectively.

As summarized in Table 2.5, the local models for Medio San Juan and Cumaral differ substantially in both selected variables and the estimated coefficients. Both models include lab_prob, population, river_length, and road_length, while coca_area, seizures, left_wing, and police are only included in the

Medio San Juan model.

In Medio San Juan, river transportation appears to be the most influential factor associated with the high predicted probability of being a cocaine destination. The coefficient of river_length exhibits a large positive value (2.49) with statistical significance, resulting in the largest positive contribution among the included variables. This finding suggests that Medio San Juan's predicted trafficking risk might be primarily linked to the accessibility of river transportation.

For context, Medio San Juan remains a critical geostrategic region for transnational criminals, serving as a vital smuggling corridor on the Pacific. The *municipio* serves as a vast hydrological basin where dozens of tributaries feed into the San Juan River, one of the largest rivers by discharge along South America's Pacific coast. The river empties into a delta approximately 60 kilometers north of the Port of Buenaventura. This intricate waterway network forms a transit route that allows trafficking operations to move illicit cargo seamlessly from inland processing zones to coastal departure points. Historically monopolized by the FARC's 30th Front, the corridor underwent severe territorial fragmentation following the 2016 peace accord (UNODC 2026). Today, the security landscape is defined by a violent, multipolar conflict as FARC dissident factions, the ELN, and the Clan del Golfo clash to secure hegemony over this fractured but vital territory (ACAPS 2021).

Cumaral, on the other hand, is predicted to be a potential trafficking node for a different set of reasons. The strongest statistically significant variables are lab_prob and road_length. In particular, road_length has the largest positive coefficient, which characterizes this area as a critical transit and micro-trafficking corridor linking the production hubs of southern Meta and Guaviare to the primary highway networks leading to Bogotá D.C. and Villavicencio. However, since Cumaral has a below-average road length (a normalized value of -0.433), this variable contributes negatively to its predicted probability despite its positive coefficient. In contrast, lab_prob has a large negative coefficient of -20.7,

indicating that a higher predicted probability of laboratory presence is associated with a lower likelihood of being a cocaine trafficking destination. Because Cumaral itself has a below-average lab_prob value, its negative normalized value (-0.41) and negative coefficient produce a strong positive contribution to the prediction of trafficking destination.

These findings highlight the distinct roles that the two *municipios* play within the broader trafficking network, carrying direct implications for antinarcotic operations. Medio San Juan functions primarily as a riverine transit hub to move illicit cargo toward the Pacific coast, whereas Cumaral operates as a critical logistics and transit region linking jungle production zones to the central highways. Because both locations present unique geographic particularities, they demand a strategic, highly localized approach rather than a generalized enforcement model. Consequently, for an inland transit corridor like Cumaral, law enforcement must concentrate on highway checkpoints, land-based interdiction, and dismantling localized supply networks; conversely, in a riverine node like Medio San Juan, the highest-impact strategies rely on fluvial interventions, mangrove-corridor interdiction, and targeted raids on riverbank storage facilities.

2.4.4 Sensitivity Analysis

Our modeling framework has two modeling parameters, n_y (the minimum number of observations required in each response class) and n_{drop} (the threshold used to exclude an independent variable). Here, we study the sensitivity of the results with respect to these two parameters. Specifically, we study how the predictive performance, as measured by AUC, changes when n_y and n_{drop} are varied from 7 to 9 and 5 to 20, respectively. We consider a relatively narrow range for n_y because larger changes can substantially alter the estimation procedure. Increasing n_y requires larger local sample

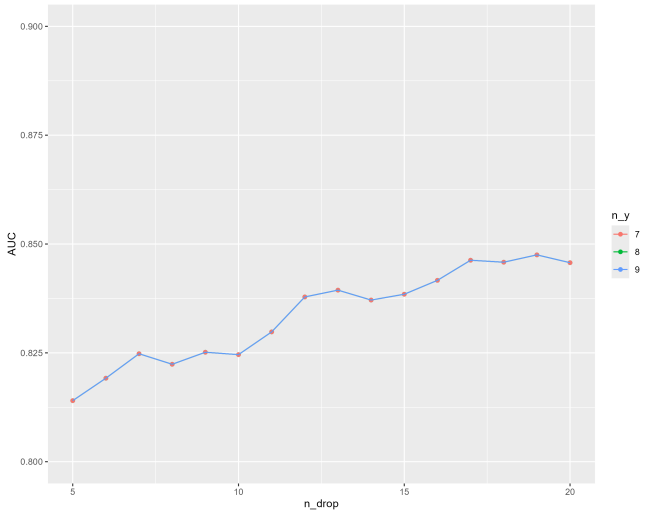


Figure 2.10: AUC of the local GWR PML model obtained with various n_y and n_{drop}. (There is no difference with varying n_y for a fixed n_{drop}. Thus, three lines with $n_y = 7$, 8, and 9 are overlapped in this figure.)

sizes to have more observations for each class in $\mathbf{y}^i$, causing more *municipios* to be included with the larger bandwidths, whereas decreasing n_y allows models to be estimated with smaller local samples. In contrast, varying n_{drop} primarily affects the set of independent variables included in each local model while leaving the best bandwidth b^* unchanged. Consequently, here we place more emphasis on n_{drop} to evaluate the robustness of the results to variable-selection decisions.

We find that the predictive performance of the local GWR PML models also remains robust across the full set of parameter values, as illustrated in Figure 2.10. The AUC values range from 0.813 to 0.848, and are not affected by n_y. The AUC exhibits a modest increasing trend as n_{drop} increases.

When selecting final parameter values, a balance must be struck between reduced model insights (due to dropping more variables) and modest improvement in model performance. We selected n_{drop} equal to 10, in part to retain the ability to evaluate the impact of coca_area, one of the central cocaine-related variables.

Table 2.6 summarizes the variation in the estimated coefficients obtained from the local GWR

Table 2.6: Sensitivity analysis for n_y and n_{drop} for each variable. (*Note*: n_{large_sd} is the number of *municipios* where the standard deviation of each coefficient is greater than 20.)

Variable	The standard deviations of local GWR PML Coefficients						
	Min	Q1	Median	Mean	Q3	Max	n_{large_sd}
coca_area	0	0.005	0.145	0.333	0.459	3.344	0
seizures	0	0.085	0.202	0.418	0.434	15.675	0
lab_prob	0	0.299	0.811	2.106	2.014	49.116	17
left_wing	0	0.122	0.309	0.436	0.571	7.618	0
right_paramilitary	0	0.048	0.242	0.455	0.559	26.376	1
left_right	0	0.016	0.164	0.404	0.486	26.278	1
population	0	0.094	0.252	0.346	0.484	3.600	0
river_length	0	0.053	0.133	0.209	0.274	2.577	0
road_length	0	0.062	0.174	0.419	0.460	9.940	0
airport	0	0.036	0.174	0.313	0.449	2.827	0
ferry	0	0.000	0.000	0.145	0.213	1.441	0
military	0	0.036	0.157	0.280	0.371	3.668	0
police	0	0.079	0.181	0.283	0.403	2.354	0

PML model under different combinations of n_y and n_{drop}. For each *municipio*, we first compute the standard deviation of each variable's estimated coefficient across 48 different combinations of n_y and n_{drop}. We then summarize these *municipio*-level standard deviations using the minimum, first quartile (Q1), median, mean, third quartile (Q3), and maximum values.

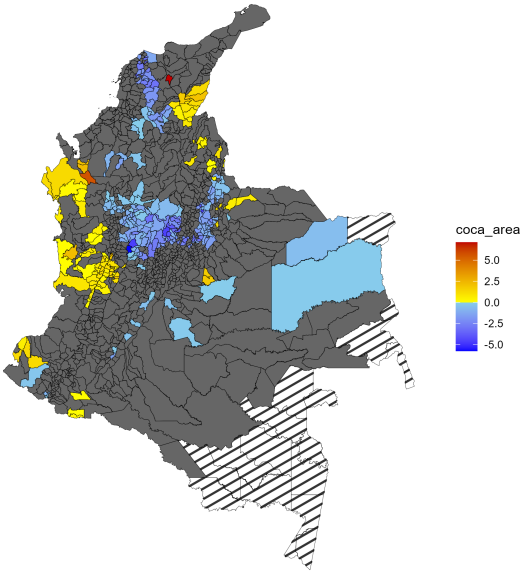
From the minimum values through the third quartiles, the results indicate that the estimated coefficients remain relatively stable across most *municipios* for all variables. In particular, the largest third-quartile value is 2.014 for lab_prob, suggesting that the coefficient variability is limited to a small range. Although lab_prob exhibits a comparatively large maximum standard deviation of 49.116, this instability is primarily attributable to partial observations of cocaine laboratories with a smaller n_{drop}, which can result in insufficient variability of lab_prob values (i.e all values are in a small range). Consequently, even when applying the PML framework, the estimation of the coefficient for lab_prob may become inflated when n_{drop} is below ten.

The column labeled n_{large_sd} in Table 2.6 reports the number of *municipios* with coefficient standard deviations exceeding 20. From this column we learn that only a small fraction of *municipios* experience substantial variation in their estimates. For lab_prob, only 17 *municipios* exceed this threshold, representing less than 2% of all *municipios* included in the analysis. These findings suggest that the proposed local GWR PML framework generally produces stable coefficient estimates across a wide range of parameter specifications (n_y and n_{drop}).

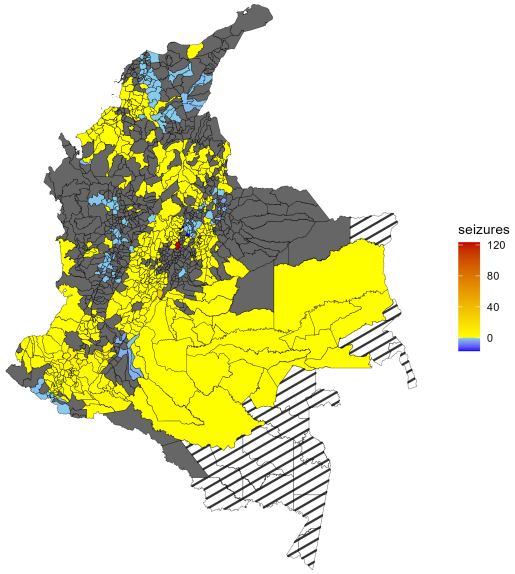
Overall, the sensitivity analysis demonstrates that the proposed local GWR PML framework yields stable and reliable regression estimates while effectively mitigating the separation and small-sample issues that commonly arise in localized binary regression models. Taken together, it supports the robustness of the proposed framework and highlights its suitability for analyzing partially observed illicit activity data characterized by spatial heterogeneity and data sparsity.

2.4.5 Comparisons to Another Approach: GWR LASSO

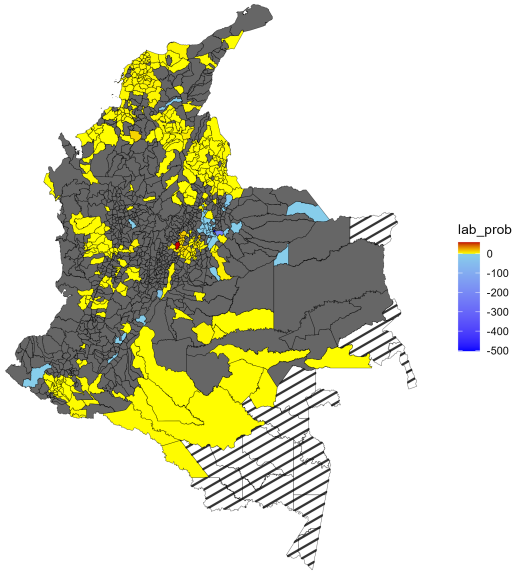
To demonstrate the advantages of the local GWR PML framework in addressing the separation issue discussed in Section 2.3.1.1, we compare the coefficient surfaces obtained from the local GWR PML (Figure 2.5) and local GWR LASSO regression models (Figure 2.11). GWR LASSO has been proposed in the literature as a means of mitigating multicollinearity in estimation (Comber and Harris 2018, Yoneoka et al. 2016). In contrast to the PML framework, the LASSO-based logistic regression does not explicitly resolve the separation problem, which can result in unstable or implausibly large coefficient estimates in regions with limited or highly imbalanced local observations. This issue is evident especially for lab_prob, where some estimated coefficients attain unrealistically extreme magnitudes, indicating overfitting (e.g., a value less than -400 in Figure 2.11(c)). Furthermore, the



(a) coca_area



(b) seizures



(c) lab_prob

Figure 2.11: Estimated regression coefficients of (a) coca_area, (b) seizures, and (c) lab_prob using local GWR lasso model.

Notes: Positive associations are represented by colors ranging from yellow to red according to their magnitude, whereas negative coefficients are shown in shades of light to dark blue. *Municipios* with coefficients equal to zero are shaded in gray, while white regions indicate locations where the corresponding variable was excluded from the local GWR model due to insufficient variation in its values among *neighbors*. Striped regions are excluded from the analysis.

GWR-LASSO results contain a considerably larger number of zero-valued coefficients, which indicates that more *municipios* exclude particular variables from their local models than the local GWR PML models.

As a consequence, the coefficient maps produced by the local GWR LASSO model appear less interpretable and exhibit weaker spatial variation than those from the local GWR PML model. These results highlight the importance of accounting for separation in localized logistic regression settings, particularly when modeling sparse and partially observed illicit activity data.

2.5 Conclusion

This study examined the spatially heterogeneous factors associated with cocaine hydrochloride trafficking destinations across Colombia. Overall, the local GWR PML framework provides a nuanced and stable geographic characterization of trafficking-related factors while effectively addressing separation issues associated with sparse binary dependent and independent variables, and avoiding overfitting and inflated regression coefficients. By producing stable estimates of spatially varying relationships, the proposed model helps narrow the range of plausible trafficking dynamics and facilitates the development of more targeted hypotheses for subsequent empirical investigation and policy analysis.

The analysis provides several substantive insights into the spatial organization of cocaine trafficking in Colombia. The key drivers of trafficking risk vary geographically: coca cultivation is most important along the Pacific coast, transportation accessibility dominates in the Andean and Amazon regions, police station presence and river access are influential in the Caribbean, and airport presence is a consistent predictor in Orinoquía.

The variation in factors associated with cocaine hydrochloride destinations suggests that trafficking

patterns cannot be explained by a single nationwide model. Seizures in prior years generally indicate persistent cocaine trafficking activity within a *municipio*, but they also show localized negative associations that may reflect displacement or deterrence in response to interdiction.

The influence of armed groups also differs geographically. Left-wing guerrilla presence is more frequently associated with trafficking destinations across Colombia than right-wing paramilitary groups. Furthermore, the interaction between the two groups varies geographically, suggesting that regional differences in competition, coexistence, or collaboration shape their influence on cocaine trafficking. Transportation infrastructure exhibits region-specific importance, with airports, rivers, roads, and ferry terminals facilitating trafficking in some regions but not others, while population is positively associated with trafficking destinations across much of the country. The case study further illustrates that *municipios* with similarly high probabilities of serving as cocaine destinations can be driven by fundamentally different local factors.

Despite these contributions, several limitations warrant consideration. First, the analysis focuses on spatial variation in trafficking destinations by pooling information in 2016 and 2017, leaving temporal dynamics unexplored. Although longitudinal data are available, annual observations of trafficking-related activities and several independent variables are considerably sparser than the aggregated dataset used in this study. As a result, extending the analysis to a spatiotemporal framework presents substantial challenges related to data availability, model stability, and local sample sizes. Future research could explore approaches for incorporating temporal dynamics while preserving the ability to estimate geographically localized relationships.

Second, as with most studies of illicit activities, the analysis relies on partially observed data that may reflect variation in detection and reporting processes. Although the predicted probability of laboratory presence helps account for potentially undetected cocaine laboratories, the effectiveness

of this approach depends on the accuracy of the lab prediction model. Incorporating additional data sources and more precise measures of laboratory and trafficking activity would strengthen the analysis and improve inference regarding the mechanisms underlying observed cocaine trafficking patterns.

Third, the GWR framework assumes that all independent variables operate at a common spatial scale, although their associations with trafficking destinations may occur over different geographic ranges. Future research could extend the model to MGWR, which estimates variable-specific bandwidths. However, adapting MGWR to binary, sparse, and partially observed data would require methods that limit instability and overfitting, such as a penalized MGWR approach.

Nevertheless, this study has two main contributions. First, it provides critical insights into the Colombian case at a historically significant moment of change in the global cocaine supply chain. Second, it develops a robust methodological framework useful in other contexts characterized by spatial heterogeneity, rare events, and partially observed data. Consequently, the proposed framework offers a methodological blueprint for future research on illicit economies, organized crime, environmental crime, and other geographically heterogeneous processes in which data limitations and rare-event outcomes pose significant analytical challenges.

Chapter 3: Improved LISA Analysis for Zero-Heavy Crack Cocaine Seizure Data to Detect Trafficking Hot Spots

The second study employs a spatial data analysis approach to identify potential trafficking hot spots using secondary observational data, such as drug seizures, in settings where no ground-truth information on trafficking routes or networks is available. A major challenge in analyzing such criminal data is their extreme sparsity and incompleteness, which often produce a large proportion of zero observations. This zero-heavy structure can substantially reduce the effectiveness of conventional spatial analysis methods.

To address this challenge, the study extends the Local Indicators of Spatial Association (LISA) framework (Anselin 1995), a widely used tool in criminology for detecting localized spatial clustering of criminal activity. Traditional LISA analysis is designed to identify areas characterized by unusually high or low levels of activity relative to their neighbors. However, when we applied the conventional LISA procedure to crack cocaine seizure data in the United States, we found that the prevalence of zeros created a critical limitation: important spatial observations could be overlooked, leading to incomplete or potentially misleading interpretations of hot spot patterns.

This study demonstrates how zero-heavy data can affect the detection of spatial clusters and examines the implications for interpreting LISA results in criminological applications. Building on this analysis, we propose an improved LISA methodology specifically designed for zero-heavy spatial data. The proposed approach provides a more appropriate representation of the underlying spatial relationships in such cases, while preserving the strengths of the original LISA framework. When applied to U.S. crack cocaine seizure data, the enhanced method identifies spatial patterns that differ

meaningfully from those detected by the conventional approach, thereby offering insights that may support more effective law-enforcement and drug-interdiction strategies.

3.1 Literature Review

Local measures of spatial autocorrelation appeared in the early 1990s as tools for detecting geographically localized patterns. Among these early contributions, the works of Getis and Ord (1992) and Anselin (1995) are particularly influential, introducing two of the most widely used LISA statistics: the G_i and G_i^* statistics, and local Moran's I, respectively. These LISA statistics enabled the identification of local hot spots, cold spots, and areas of spatial mismatch—locations where a unit's local spatial clustering significantly deviates from the global average. Since their introduction, LISA statistics have been broadly applied in criminology literature, including studies of the spatial impact of murders (Cohen and Tita 1999) and terrorist activity (LaFree et al. 2012).

According to Anselin (1995), the LISA statistics are defined as follows:

Definition 3.1.1. A local indicator of spatial association (LISA) is any statistic that satisfies the following two requirements:

1. The LISA for each observation gives an indication of the extent of significant spatial clustering of similar values around that observation.

2. The sum of the LISAs for all observations is proportional to a global indicator of spatial association.

Unlike global measures of spatial autocorrelation, LISA statistics identify specific locations that exhibit significant spatial correlation with their neighbors. These statistics provide localized insights

into spatial clustering patterns that global indicators may overlook. When the data distribution is known, statistical significance can be evaluated using conventional tests. Otherwise, significance can be assessed using a permutation-based approach, in which the observed data serve as the reference distribution to compute pseudo p-values (Anselin 1995). This method is commonly referred to as conditional randomization or permutation inference.

Recently, LISA analysis has been extended in several ways. Some of them tried to expand the formula of LISA statistics from univariate to multivariate. For example, Anselin (2019) introduced a multivariate version of Geary's c, which can measure feature similarity and locational similarity at the same time. Liu et al. (2015) developed spatial autocorrelation measures for vector data that incorporate both magnitude and direction, making them well-suited for analyzing "flow" data such as wind patterns, the distribution of products, migration, communication, and capital movement. Another line of research has expanded LISA to longitudinal spatial data, building on earlier efforts to incorporate temporal dynamics in the data. While prior approaches incorporated local temporal dynamics, they failed to adjust for evolving global temporal trends because local Moran's I is defined relative to the global average. To address this, Tao and Chen (2023) proposed an "absolute perspective" version of local Moran's I that accounts for temporal trends in the data. Despite these advances, none of these studies have addressed the challenges posed by zero-heavy data in LISA analysis.

The most closely related works were about regression models that used Poisson or zero-inflated Poisson distribution to incorporate rare response variables. Feng (2021), for example, introduced hurdle models with two-stage regression models: one for the likelihood of a non-zero outcome, and another for the magnitude of non-zero values. With longitudinal data, regression models often incorporate spatial autocorrelation through spatially structured random effects, typically assumed to be normally distributed. Applications include studies of offender counts in Sheffield, England (Haining et al. 2009),

and lung cancer incidence in Ontario (Feng and Dean 2012). To the best of our knowledge, however, no prior work has adapted LISA analysis for zero-heavy data. This chapter fills this gap by demonstrating the limitations of applying conventional LISA techniques to such data and proposing methodological improvements to enhance their applicability.

3.2 Data Overview

At first, this chapter assumes that the amount of interdicted illegal drugs, or "seizure", is proportional to the actual trafficking activities. Under the assumption, LISA analysis is implemented to crack cocaine seizure data in the U.S. collected by the High Intensity Drug Trafficking Areas (HIDTA) program.[1] The HIDTA program aims to disrupt illicit drug trafficking by cooperating with U.S. federal, state, and local governments and tribal law enforcement agencies. Their tasks include collecting and sharing drug seizure data (Washington/Baltimore High Intensity Drug Trafficking Area 2022). This HIDTA dataset contains four years (2018–2021) of drug seizure records. Each entry includes the county name, the name of the seized drug, its weight in kg, and the date of the seizure. In this study, it is assumed that the seizure amount is 0 for a given period if no seizure is reported for the period in a county participating in HIDTA. On the other hand, seizures for non-participating HIDTA counties are assumed to be missing, unless data are reported. The 48 mainland states (Alaska and Hawaii are not considered) and the District of Columbia are included in this study. Of 3,108 counties in total, 1,518 are HIDTA counties or have non-zero seizures. Counties with missing data on seizures are excluded from the study. As a result, some counties are abutted or sometimes fully surrounded by excluded counties. This makes LISA analysis inapplicable since it works when all the neighbors of a focal point have

[1] This program is administered by the Office of National Drug Control Policy. Visit `https://www.hidtaprogram.org` for more details.

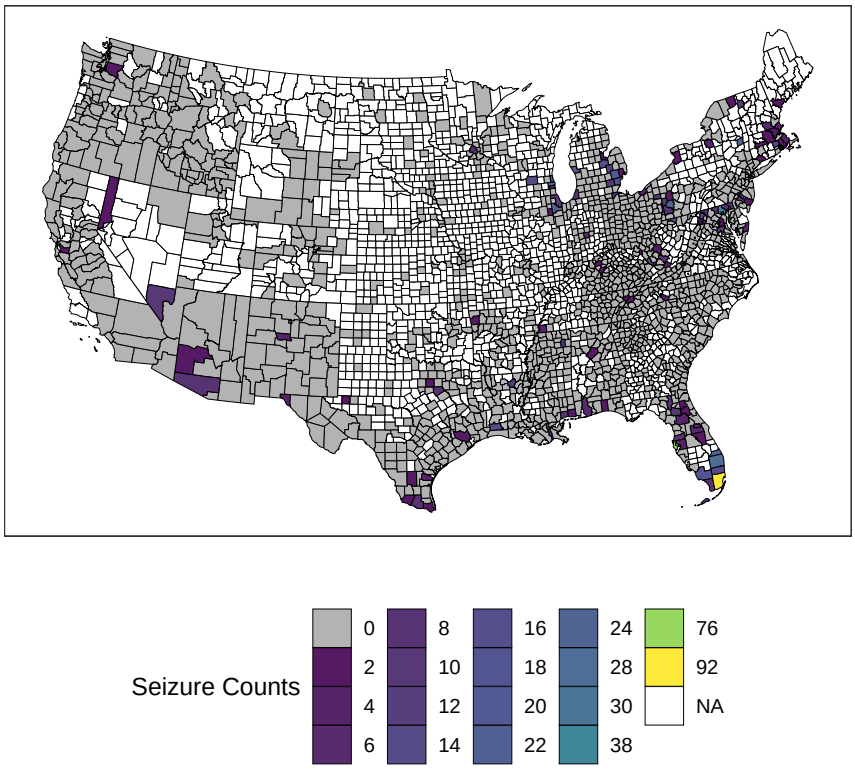


Figure 3.1: The number of crack cocaine seizures in January 2020 by county.

observations. To circumvent this issue, the proposed approach does not use a geographic neighborhood. Instead, it assigns each county the k closest reporting counties as its neighbors.[2] The value of k is fixed at 5 for all counties in the analyses that follow throughout the chapter. This value is selected based on the average number of geographical neighbors, 5.8, for the 3,108 U.S. mainland counties. It is rounded down to reduce the number of counties with "too far away" neighbors.

In this study, the focus is on crack cocaine data because of its distinct geographic patterns; the 2020 National Drug Threat Assessment (Drug Enforcement Administration, U.S. Department of Justice 2020) reported that crack cocaine is primarily produced and distributed by U.S. domestic criminal

[2]See `https://geodacenter.github.io/workbook/4b_dist_weights/lab4b.html` for more details on implementing the k closest neighbors approach.

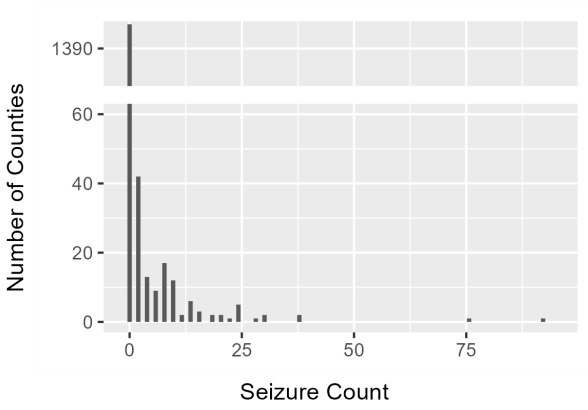


Figure 3.2: Histogram of county-level crack cocaine seizure counts in January 2020.

groups, unlike other types of cocaine. Finally, seizure data is used after being aggregated into the monthly seizure *counts* with the assumption that the frequency of trafficking uncovered is more related to the intensity of crime activities than the amount of drugs interdicted.

As an example, Figure 3.1 shows the number of crack cocaine seizures in January 2020, represented by a color spectrum, for each county. Seizure-missing counties are colored in white. Only a few counties have high seizure counts; the two highest counts are 92 for Miami-Dade, FL and 76 for Pinellas, FL. Figure 3.2 is the histogram of seizure counts shown in Figure 3.1. It indicates approximately 92% of counties have zero seizures, illustrating the zero-heavy non-normal distribution of the spatial data of crack cocaine seizure counts. The most notable states with the highest seizure counts are Florida (314 seizures) and Maryland (132 seizures). However, it is hard to gauge the intensity of spatial association based solely on this raw seizure data at first glance. In this case, LISA analysis can be useful in identifying areas of local patterns, hot spots, cold spots, and spatial mismatches. When applied, it can provide valuable insights to law enforcement and policymakers, for instance, by indicating warranted areas with heightened attention when faced with resource constraints.

3.3 Local Indicator of Spatial Association (LISA) Statistic: Local Moran's I

Anselin (1995) explained the main advantage of LISA statistics over global indicator. First, LISA can assess the statistical significance of local spatial autocorrelation or clustering. Therefore, LISA statistics can highlight spatial association outliers, or some localized areas that are different from their surrounding neighbors. In other words, it provides us with local-level information about spatial correlation around each location, which helps identify local areas of interest that can be missed using global statistics.

Formally, a LISA statistic can be written as $L_i = f(x_i, x_{J_i})$, where x_i is a value at location i and x_{J_i} are the values observed in the neighborhood J_i of i. For example, the standardized local Moran's I (I_i) is defined by Anselin (1995) as:

$$I_i = \frac{\sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{j=1}^{N}(x_j - \bar{x})^2/N}, \tag{3.1}$$

where N is the number of locations, $\bar{x}$ is the sample mean of x, and w_{ij} are weight coefficients with $w_{ij} \geq 0$ if j is a neighbor of i and $w_{ij} = 0$ otherwise or $i = j$. Local Moran's I at i, I_i, represents the spatial autocorrelation among i and its neighbors.

Practically, geographers who utilize LISA analysis generally translate the values of I_i into five categories, which we refer to as *LISA labels*: *Insignificant* (*Insig*), *High-High* (*HH*), *Low-Low* (*LL*), *High-Low* (*HL*) and *Low-High* (*LH*). These labels indicate whether the spatial autocorrelation surrounding the focal point i is statistically significant and the direction of the autocorrelation in both the focal location i (the first or left letter) and its neighbors (the second or right letter).

3.4 Statistical Significance Test of LISA Statistics: Permutation Inference

Moran's I calculated from a dataset do not necessarily follow a well-defined distribution like a normal distribution. In this case, the statistical significance of I_i is typically tested with nonparametric, empirical inference, known as permutation inference (see Anselin 1995, for details). In summary, permutation inference uses the simulated data drawn multiple times from the empirical distribution of observed values x_i. The standard approach of permutation inference is to generate a specific number M (typical values are 999, 9999, or 99999) of permuted Moran's I for I_i, or I_i^m where $m \in \{1, 2, ..., M\}$, with a fixed local value x_i using spatial shuffling.

Formally, the m^{th} permuted I_i is defined as:

$$I_i^m = \frac{\sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_{(j)}^m - \bar{x})}{\sum_{j=1}^{N}(x_j - \bar{x})^2/N}, \tag{3.2}$$

where (j) is the index of permuted neighbor x_j. The only difference from I_i (Equation 3.1) is that x_j is now $x_{(j)}^m$, which refers to a neighbor's observed value sampled from the set $\{x_1, ..., x_N\}$. Then, a set of M values of I_i^ms forms a reference distribution for I_i. This empirical distribution is used to derive a pseudo p-value for I_i, which allows us to determine the significance of I_i. Specifically:

$$p = \frac{R + 1}{M + 1}, \tag{3.3}$$

where R is the number of times the permuted Moran's I is equal to or more extreme than the observed Moran's I. R is defined as $R^+ := \sum_{m=1}^{M} I\{I_i^m \geq I_i\}$ for upper-tail tests and $\min\{R^+, M - R^+\}$ for two-sided tests (this follows the implementation of the R package "spdep"[3] as in Hughes and MacKenzie

[3] `https://www.rdocumentation.org/packages/spdep/versions/1.2-8`

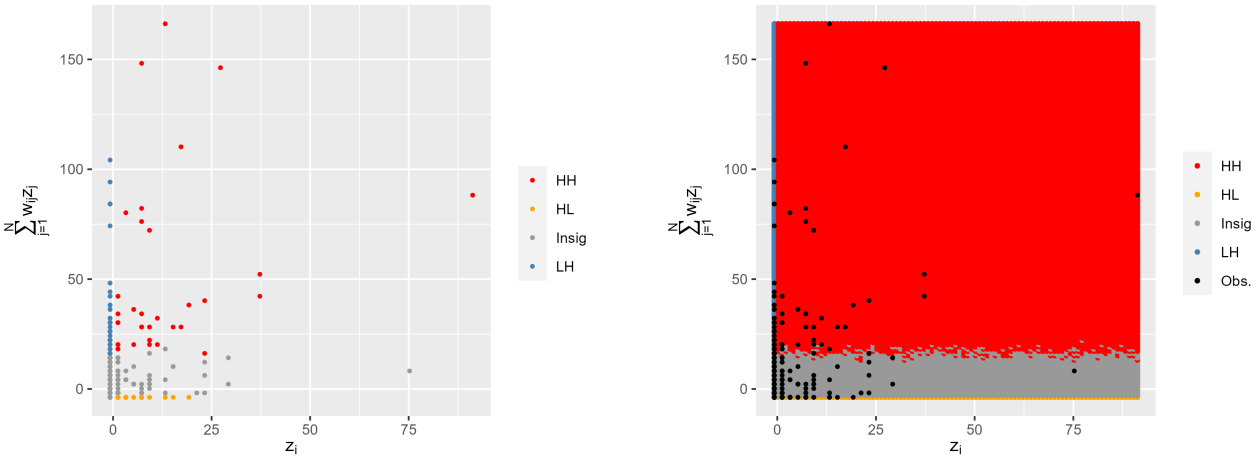
x_i	Observed value at location i
$\bar{x}$	The sample mean of x_i
z_i	$x_i - \bar{x}$
$x_{(i)}$, $z_{(i)}$	ith element in the permuted sequence of x_i and z_i
I_i	Local Moran's I at location i
I_i^m	mth permuted local Moran's I at location i for $m \in [1, M]$
M	The number of permutations taken in permutation inference
N	The total number of locations
w_{ij}	A weight representing the neighborhood between two locations i and j with $w_{ii} = 0$
J_i	A set of $j \in \{1, ..., N\}$ where $w_{ij} = 1$ at location i
S_{J_i}	$\sum_{j \in J_i} x_j$

Table 3.1: Notation for Section 3.3 and 3.4.

(2016) and Oldekop et al. (2019)). If the pseudo p-value is smaller than a significance level $\alpha \in [0, 1]$, I_i is said to be statistically significant. Note that the significance is independent of the value of x_i and is determined solely by the value of S_{J_i} since x_i stays fixed and is canceled out in the inequality when the M different I_i^m are generated. Henceforth, the process described above will be referred to as "standard" method. We summarize the notations of variables in Table 3.1.

3.5 Limitations in LISA Analysis with the Standard Method

Conventionally, LISA analysis visualizes the local Moran's I as a scatter plot, known as Moran's scatter plot, with respect to the focal unit (x_i) and its surrounding units $\left(\sum_{j=1}^{N} w_{ij} x_j\right)$ to easily understand the characteristics of spatial autocorrelation of a given data in one figure. Then, the LISA labels, which indicate the significance and direction of I_i, can be applied to the plot with colors as shown in Figure 3.3(a). Moreover, it is possible to visualize LISA labels that would be obtained for the other unobserved values using the same empirical distribution with the actual observations colored black, as Figure 3.3(b), which we would call the "significance region". Note that the borders between significant (red and orange) and insignificant (gray) regions are horizontal. This is because the significance of I_i is



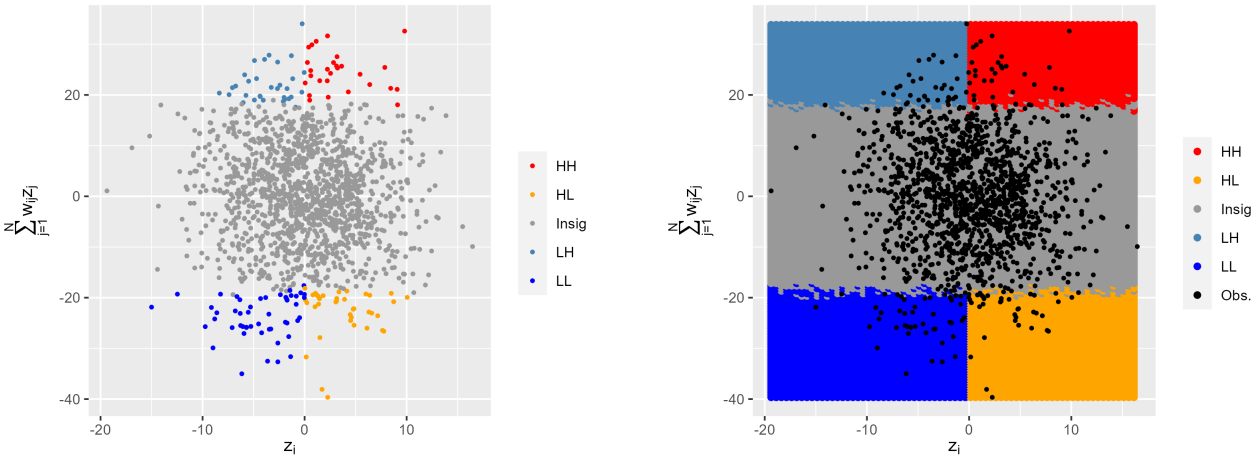
(a) **(b)**

Figure 3.3: Moran's scatter plots of (a) centered crack cocaine seizure counts ($z_i = x_i - \bar{x}$) in January 2020 versus the sum of its neighbors' seizure counts $\left(\sum_{j=1}^{N} w_{ij}z_j\right)$ and (b) its significance region with $M = 999$, $k = 5$, and $\alpha = 0.1$.

independent of the value of z_i as mentioned in Section 3.3.

Figure 3.3 highlights that the standard LISA analysis fails to identify certain counties with high levels of trafficking activity. For example, in the lower right of Figure 3.3(b), Pinellas County, FL —despite exhibiting very high seizure counts— is labeled as insignificant since neighbors with low positive values surround it. From a drug enforcement perspective, it is counterintuitive to exclude such a county from further investigation for that reason. This illustrates that relying solely on standard LISA analysis may lead to overlooking regions that are actively involved in cocaine trafficking.

For comparison, we generated a normally distributed sample x_i with a mean of 10 and a variance of 25. Figure 3.4 is Moran's scatter plot and the significance region with the normally distributed sample. Unlike Figure 3.3, LISA labels with the normally distributed data exhibit a symmetric configuration of the significance region. We also observe horizontal boundaries between significant and insignificant regions, but almost equally stretched toward both positive and negative directions on the y-axis. This is because of the symmetry of the normally distributed sample, where they are equally distributed from its center at $(0, 0)$.



(a) **(b)**

Figure 3.4: Moran's scatter plots of (a) centered values z_i of normally distributed random samples $x_i \sim N(10, 25)$ with sample size of 1,518 and (b) its significance region with $M = 999$, $k = 5$, and $\alpha = 0.1$.

One key difference between Figure 3.3 and 3.4 is that the LL label does not appear with zero-heavy data. We can prove that none of the samples will be labeled as LL when the proportion of zero-valued observations exceeds a certain point:

Proposition 3.1. Let x_i be a non-negative random variable for $i \in \{1, 2, ..., N\}$, and let N_0 denote the number of zero-valued observations, or $N_0 = \sum_{i=1}^{N} I(x_i = 0)$. Assume that each of the N focal points has k neighbors. If $\frac{\binom{N_0}{k}}{\binom{N}{k}} > \alpha/2$, where α is the significance level for the permutation inference, then no focal point will be labeled as LL with the standard approach.

Proof. To demonstrate Proposition 3.1, we note that for a location to be labeled as LL, we need $z_i < 0$, or $x_i < \bar{x}$, and $\min\{P(I_i \leq I_i^m), 1 - P(I_i \leq I_i^m)\}$ needs to be smaller than $\alpha/2$ (for a two-sided test). Since we are discussing the LL label, we will assume that $x_i < \bar{x}$ and show that even with the lowest possible values for the neighbors, all of them zero, $\min\{P(I_i \leq I_i^m), 1 - P(I_i \leq I_i^m)\} > \alpha/2$. Recall that

$$P(I_i \leq I_i^m) = P\left((x_i - \bar{x}) \sum_{j=1}^{N} w_{ij}(x_j - \bar{x}) \leq (x_i - \bar{x}) \sum_{j=1}^{N} w_{ij}(x_{(j)} - \bar{x})\right).$$

Since $x_i < \bar{x}$, the inequality in the probability is reversed when $(x_i - \bar{x})$ is canceled out, and therefore

$$\begin{aligned} P(I_i \leq I_i^m) &= P\left(\sum_{j=1}^{N} w_{ij}(x_j - \bar{x}) \geq \sum_{j=1}^{N} w_{ij}(x_{(j)} - \bar{x})\right) \\ &= P\left(\sum_{j=1}^{N} w_{ij}x_j \geq \sum_{j=1}^{N} w_{ij}x_{(j)}\right). \end{aligned}$$

With all neighbors assumed to be zero, $\sum_{j=1}^{N} w_{ij}x_j = 0$. Then, we have:

$$\begin{aligned} P(I_i \leq I_i^m) &= P\left(0 \geq \sum_{j=1}^{N} w_{ij}x_{(j)}\right) \\ &= P\left(0 = \sum_{j=1}^{N} w_{ij}x_{(j)}\right). \end{aligned}$$

Now, with the number of zero-valued observations N_0, the probability of all k neighbors having zero values[4] ($\sum_{j=1}^{N} w_{ij}x_{(j)} = 0$) is

$$P\left(\sum_{j=1}^{N} w_{ij}x_{(j)} = 0\right) = \frac{\binom{N_0}{k}}{\binom{N}{k}}$$

As long as this probability is greater than $\alpha/2$, we have:

$$P(I_i \leq I_i^m) = P\left(0 = \sum_{j=1}^{N} w_{ij}x_{(j)}\right) > \alpha/2. \quad \blacksquare$$

Based on Proposition 3.1, when one chooses the significance level $\alpha = 0.05$, for example, with five neighbors for all focal points, there will be no LL labels even with a mild zero-heaviness of

[4]Recall that permuted neighbors $x_{(j)}$ are selected without replacement

$N_0/N = 0.5$ for $N = 1518$, as

$$P\left(0 = \sum_{j=1}^{N} w_{ij}x_{(j)}\right) = \frac{\binom{759}{5}}{\binom{1518}{5}} \approx 0.03104 > \alpha/2 = 0.025.$$

Our seizure data has many zero-valued locations, with $N_0/N = 0.92$ and $N_0 = 1397$, which demonstrates why no LL label is observed in Figure 3.3 since

$$\frac{\binom{1397}{5}}{\binom{1518}{5}} \approx 0.6597.$$

3.6 Improved LISA Analysis

As discussed in Section 3.5, the standard LISA significance test depends solely on the values of neighboring observations because the focal observation remains fixed during permutation inference. This limitation becomes particularly problematic when the data are highly skewed and contain a large proportion of zero observations. In such settings, locations with unusually large values may be classified as insignificant simply because they are surrounded by neighbors with relatively low values. Motivated by these limitations, we propose two extensions to the conventional LISA framework: first the permuted i method and second the moderate label method, as well as a combined approach that integrates both modifications.

3.6.1 Permuted i Method

In the permuted i method, the significance of LISA statistics is evaluated based on observations from both a focal location and its neighbors. The updated LISA analysis has the following modified

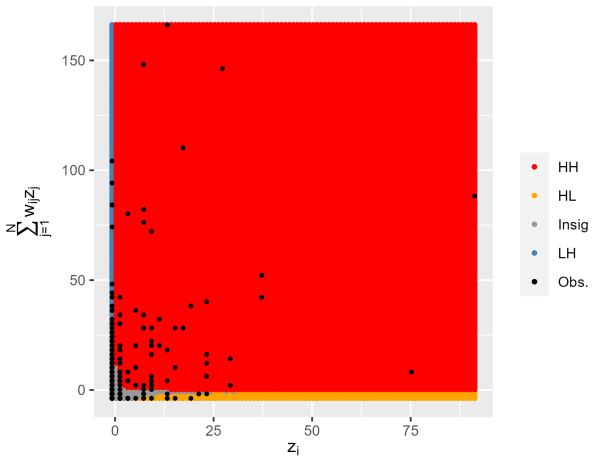


Figure 3.5: Significance regions using the permuted i approach for centered crack cocaine seizure counts z_i vs. sum of neighbors' seizure counts $\left(\sum_{j=1}^{N} w_{ij}z_j\right)$ with $M = 999$, $k = 5$, and $\alpha = 0.1$.

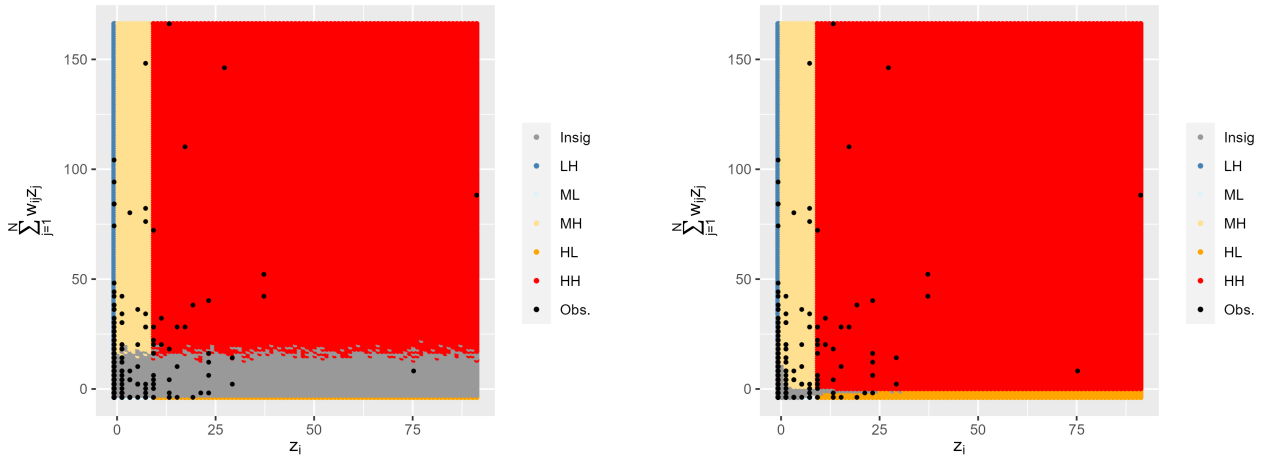
version of the permuted local Moran's I formula:

$$I_i^m = \frac{\sum_{j=1}^{N} w_{ij}(x_{(i)} - \bar{x})(x_{(j)} - \bar{x})}{\sum_{j=1}^{N}(x_{(j)} - \bar{x})^2/N}, \tag{3.4}$$

where $x_{(i)}$ is the permuted observation for a local point i. The difference from the standard method (Equation 3.2) is that, in Equation 3.4, x_i is permuted as well as the values of neighbors $x_{(j)}$. As a result, the value of a focal point, x_i, also contributes to the significance of I_i. Then a county with a relatively high seizure is more likely to be significant, even when its neighbors have relatively lower seizures. This property is especially important when working with zero-heavy data, as larger positive observations are considered rarer events and are therefore regarded as more valuable for further investigation.

3.6.2 Moderate Label and Combined Methods

Motivated by the need to distinguish between moderately high and very high values in a focal county, we propose the use of a "moderate label", M (Getis and Ord 1992), which adds detail to our



(a) **(b)**

Figure 3.6: Significance region with the zero-heavy data using the moderate label with (a) the standard approach and (b) the permuted i approach with $M = 999$, $k = 5$, and $\alpha = 0.1$.

interpretation of the results. The M label enables us to distinguish moderately positive values from very high ones, which is especially important for zero-heavy data as the range of values indicating a significant region of HH is typically very large. Adding the M label allows the analyst to compare the intensity of the significant values, which can be vastly different. However, introducing a new label adds complexity. One additional label for local values increases the number of LISA labels from four (HH, HL, LH, and LL) to six (adding MH and ML). If we add a "moderate" label for the neighbors as well, then there are three additional LISA labels (HM, LM, MM).

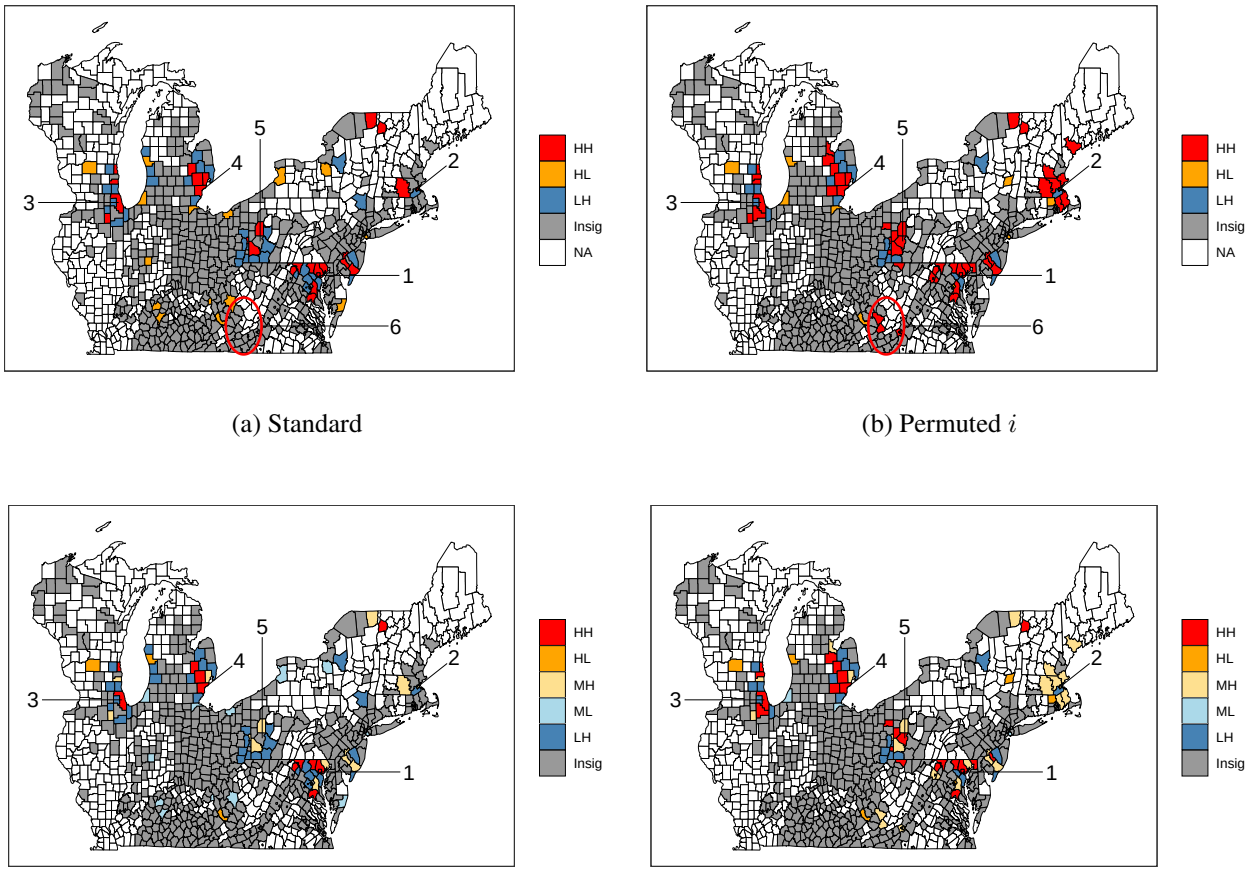
The threshold for the new M label should be selected to align with the interpretation of what qualifies as moderate or extreme within the specific context of the application, based on the data distribution. For example, in our crack cocaine seizure data displayed in Figure 3.6, we decided to add the M label to the focal county only and set the boundary corresponding to $\bar{x}$ plus two standard deviations of seizure counts; seizures beyond this threshold signify high activity. This label will aid in the interpretation of the resulting analysis, drawing differences between seizure activity in the focal location. We then combine the permuted i approach and add the moderate label, referring to these two

updates as the "combined method".

3.7 Case Study with the Northeastern States and Florida

To highlight the potential of the improved LISA analysis, we compare the LISA labels results to the standard method using crack cocaine seizure data in the Northeastern states. We visualize the results of the four methods in Figure 3.7, which highlights five major cities (Baltimore, Boston, Chicago, Detroit, and Pittsburgh). In Figure 3.7(a), displaying the standard LISA approach, the labeled counties each exhibit a notably high or low local Moran's I. Prominent HH clusters are identified around the five cities, while several HL counties are observed away from these urban areas. These results serve as a reference against the permuted i method, contrasting the different insights they possess.

Some previously significant counties in the standard approach (Figure 3.7(a)) become insignificant in the permuted i (Figure 3.7(b)) approach. There are two reasons for this transition: i) the seizure count of the focal county is less than five (HL to Insig), reflecting a low seizure count in the focal county and zero seizures among its neighbors, or ii) neighbors' seizure counts are low (LH to Insig). By not labeling these relatively low-seizure areas as areas of importance, the permuted i method allows the analyst's focus to move to higher activity areas. The new criteria for the significance test of permuted i intensifies existing HH clusters by increasing the number of HH counties around the five cities and reveals additional HH clusters, e.g., Portland, Maine (north of Boston), and an area on the border between West Virginia and Virginia (identified as number 6 in Figure 3.7(b)). Although these areas may merit the interest of law enforcement and/or policymakers due to their high seizure counts, they were labeled insignificant by the standard method because they are surrounded by neighbors with moderate seizure counts. This reflects our earlier discussion of the differences in significance regions identified



(a) Standard

(b) Permuted i

(c) Moderate Label

(d) Combined

Figure 3.7: The Northeastern states' LISA labels for crack cocaine seizure data from January 2020 under (a) standard, (b) permuted i, (c) moderate label, and (d) combined (key cities are labeled as 1: Baltimore, 2: Boston, 3: Chicago, 4: Detroit, 5: Pittsburgh, and 6: the border between Virginia and West Virginia).

by the different approaches.

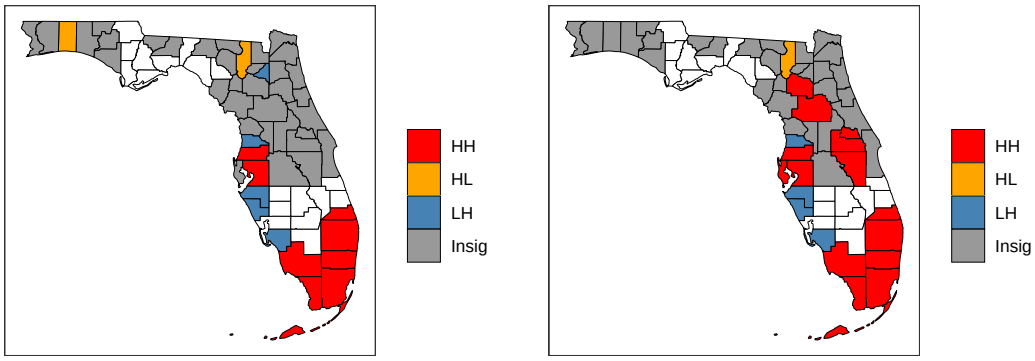
The moderate label approach introduces additional information about the intensity of seizure counts for individual focal points. The significant counties obtained by the moderate label method remain consistent with those identified by the standard method, and the only change is that certain counties previously labeled as HH or HL are now labeled MH and ML, respectively. This is illustrated in Figure 3.7(c). By introducing the distinction between HH and MH, we gain the ability to evaluate the strength of the focal county's activity within a cluster. For example, from Figure 3.7(c), we can

infer that the cluster around Pittsburgh (MH cluster) exhibits a weaker intensity than the cluster around Baltimore (HH cluster). This allows for a more detailed interpretation of the areas of interest. The moderate label method introduces an additional dimension to the HL classifications as well: counties with moderately positive seizure counts surrounded by zero-count neighbors, like Berrien, Michigan (a county east of Chicago), are now labeled ML, and counties with extremely high seizure counts also surrounded by zero neighbors like Muskegon, Michigan (a county in west Michigan) are labeled HL.

Similar patterns are observed when we compare the results from the permuted i and combined approaches. The significant counties in Figure 3.7(d) are consistent with those in Figure 3.7(b). However, some counties previously labeled HH and HL are now labeled MH and ML, respectively. The introduction of the new M label again provides insight into the level of seizure activity; it helps us identify whether the newly significant counties (when compared to the standard method) exhibit high (HH and HL) or moderate (MH and ML) seizure counts. For example, when comparing the intensity of spatial autocorrelation between and within HH clusters, we note that Boston demonstrates relatively weaker intensity and lower seizure activity than other HH clusters. Also, within HH clusters, some counties in the Baltimore, Chicago, Detroit, and Pittsburgh clusters exhibit lower seizure counts than the other counties in their HH clusters. Collectively, the results from the improved LISA approach focus the analyst's attention on areas of higher importance and provide a more nuanced view of the data, therefore better supporting interdiction strategies by law enforcement and insights for policymakers.

We next study the labels in Florida resulting from the standard and the permuted i methods. This serves to demonstrate an additional advantage of the permuted i method and underscores the rationale for its consideration. Figure 3.8 summarizes the results.

When comparing the two approaches we note that in Figure 3.8(a), the majority of HH counties are concentrated in the southern region of Florida. Additionally, there are two HH counties in the



(a) Standard (b) Permuted i

Figure 3.8: Florida's LISA labels for crack cocaine seizure data in January 2020 under (a) standard and (b) permuted i approaches.

middle-western area and two HL counties along the northern border. Given these results, one could hypothesize that the most likely trafficking route for crack cocaine originates in the southern counties and moves northwards along the western coastline of Florida. However, the presence of additional HH counties in Figure 3.8(b) indicates the high activity —and the potential for major trafficking routes— along key highways traversing central Florida. In this case, the differences between the standard and the permuted i methodologies directly impact the interpretation of the drug seizure data and can consequently point to the need for different interception strategies. This illustrates how the different approaches lead to different insights and interpretations, and we argue that the permuted i approach is better suited for highlighting relevant activity when the data is zero-heavy.

3.8 Conclusion

In this study, we propose an improved LISA method and a more detailed labeling approach that allows analysts to detect trafficking hot spots more reliably in the presence of zero-heavy data, a common feature in many types of illicit activity reporting. Our method will provide better decision-

making support for geographic data contexts in diverse fields, including but not limited to criminology or business problems.

Although the proposed methodology improves LISA analysis for zero-heavy data, several limitations remain. The approach inherits many assumptions of the underlying LISA framework, including the need to specify an appropriate neighborhood structure and the reliance on a numerical outcome variable. Moreover, the method is designed for cross-sectional spatial data and does not explicitly account for temporal dependence. Future research could extend the proposed framework to spatiotemporal settings, allowing analysts to examine how localized spatial associations evolve over time. Such extensions would further enhance the usefulness of LISA-based methods for studying illicit activities and other phenomena characterized by sparse and highly skewed spatial data. Given the critical role that LISA plays in policy analysis, this study serves to advance LISA analysis by appropriately accounting for zero-heavy property in data.

Chapter 4: Data-Driven Optimization Model for Cocaine Flow Networks Inference

The third study develops a novel data-driven optimization approach to maximize the potential information gained from noisy and inconsistent data to detect illicit networks. Using real-world data on state-level cocaine seizures and undercover purchases as well as records on overdose deaths in the U.S., the approach derives network structures that best align with available data and validates them with the DEA report on U.S. cocaine trafficking. By varying model parameters, the model creates multiple networks highlighting different plausible network structures, thereby offering unique insights and demonstrating the utility of the approach. The framework has the potential to become a tool for law enforcement agents and policymakers to guide strategies and policies to combat illicit activities.

4.1 Literature Review

Researchers have developed a number of approaches to predict or discover networks. We will review key approaches below, and further detail why they are not appropriate in our setting.

Link prediction or network completion models aim to learn or derive unknown network structures and links. This includes both similarity-based algorithms (e.g., Holme and Huss 2005) and different maximum likelihood methods (Clauset et al. 2008, Doreian et al. 2005). We refer the reader to Lü and Zhou (2011) for a detailed overview. However, link prediction models require at least some part of the network structure to be known (typically with incomplete network data or via assumptions on the node connections). In our case, we do not have *any* of the network structure; our approach necessarily does not assume an underlying network structure, because no such information is available. This also makes the related idea of structural estimation modeling problematic in our setting. Although

researchers are most likely familiar with structural estimation in the context of economics or finance (e.g., Li et al. 2002, Strebulaev and Whited 2012), structural estimation has been recently adopted in fields ranging from management to sociology (e.g., Kolenikov 2011, Musalem et al. 2010) to address questions where an underlying model is unknown but can be reasonably assumed from similar problems or other available information. However, because structural estimation naturally requires researchers to assume an underlying model specification, structural modeling techniques become increasingly biased as models are misspecified (Kaplan 1988, Kolenikov 2011, Yang and Green 2010). Thus, structural estimation works best when there is a strong, if ultimately unknown, model basis. Since no part of the illicit drug supply network is either approximately known or can be concretely inferred, structural modeling is at best inappropriate and at worst misleading. Other more complex machine learning approaches are similarly not appropriate in our setting, for example dynamic causal modeling (Friston et al. 2011), as here again, the dynamic network is estimated using structural estimation.

The key contribution of this study is to introduce a new approach to learn networks in the absence of any labeled data and/or when traditional machine learning approaches do not apply. We build our ideas on the illicit network literature, which has used price correlations in illicit drug data across pairs of locations to infer potential trafficking flows. Benítez et al. (2019), for example, inferred cocaine trafficking networks in Colombia using correlations in cocaine prices. More recently, Chandra et al. (2025) incorporated cocaine purity alongside price to estimate cocaine flows across the United States. However, neither study used an optimization framework to directly infer the structure of the trafficking network. Bjarnadottir et al. (2024) developed an optimization model related to cocaine and methamphetamine flows in the United States, but the model was designed to identify an appropriate time window for aggregating the underlying data. In contrast, the optimization framework proposed in this study is used to directly infer plausible cocaine flows across the United States. We combine

multiple noisy, incomplete data sources and use optimization to combine the data sources to discover the underlying *unknown* illicit network. In doing so, we provide a novel modeling technique for estimating network flows. We show not only the empirical strength of our technique but also compare our findings to available anecdotal evidence. As further validation, we explore constraint relaxations and suggest other sensitivity analyses for understanding possible variations of the unknown network. These verification steps are important for networks where there is no ultimate ground truth available.

4.2 Data Overview

To demonstrate our approach, we use multiple data elements from two data sources. The first data set is comprehensive data of undercover purchases and seizures from the System to Retrieve Information on Drug Evidence (STRIDE). The STRIDE data are collected by the U.S. Drug Enforcement Administration (DEA) and contain information on illegal drug seizures and undercover purchases by its agents. The suitability of STRIDE data for policy analysis and research has been questioned in the literature. In particular, STRIDE was collected primarily to support criminal investigations rather than to provide a representative measure of illicit drug market prices (Horowitz 2001). As a result, the data is noisy and biased to differences across data-collection agencies, measurement error, or temporal inconsistencies. However, research has demonstrated their strengths in both policy analysis and network detection (Arkes et al. 2008, Bjarnadottir et al. 2024, Chandra et al. 2025).

We restrict the STRIDE data to observations from 2009 to 2014 (the most recent observations we have) and include all 48 contiguous U.S. states and the District of Columbia (thus excluding Alaska and Hawaii), resulting in 11,692 observations. The data include the month, state, drug type, weight, purity, and the price (in the case of a purchase) of each seizure or purchase. To develop a robust price

N	The number of states
s_i	Cocaine inflow from outside the U.S. into state i
t_i	An indicator variable indicating inflow from outside the U.S. $I(s_i > 0)$
w_{ij}	Cocaine flow from adjacent state i to adjacent state j
o_i	Overdose deaths in state i
c_i	The maximum cocaine seizure in state i
p_i	Median purity-adjusted price in state i

Table 4.1: Decision variables and key parameters.

estimate, we restrict the weight to be between 5 g and 1 kg, which reflects the "middle market" (with smaller price variance compared to the retail market), resulting in 5,583 observations. Further, the price of cocaine varies with its purity, we therefore calculate the "purity-adjusted price," which reflects the price of a gram of fully pure cocaine:

$$\text{purity-adjusted price} = \frac{\text{price}}{\text{cocaine weight} \times \text{purity}}.$$

The second data set we use is fatal overdose information collected by the Centers for Disease Control and Prevention (CDC)[1]. Consistent with CDC definitions Ahmad et al. (2024), the count of overdose deaths was extracted from the CDC website using a key ICD-10 code: T40.5 (Poisoning by cocaine). During our time period, the number of reported overdose deaths ranged between 7,500 and 11,000 deaths annually.

4.3 Network Optimization Framework

We represent each state and the District of Columbia as a node indexed with i. The aim is to learn the inflow to each state (s_i) and the flow from i to another adjacent state j (w_{ij}). Table 4.1 summarizes the key notation used in the optimization approach.

[1] `https://wonder.cdc.gov/mcd.html`

The modeling approach relies on three key assumptions. First, cocaine consumption is proportional to overdose deaths. Second, cocaine flow increases its price (due to transport costs and risk). Lastly, maximum seizure at a state is assumed to be indicative of the underlying cocaine flow in a state. Note that the absence of large seizures is not necessarily indicative of no cocaine flow - it could be a result of limited enforcement resources, for example. However, large seizures cannot happen without large flows.

Since the data are inconsistent and noisy, imposing these assumptions as strict constraints can produce infeasible or unrealistic solutions. We therefore develop three model variants with different relaxation mechanisms. The first model relaxes the seizure constraint (nationally). The second model removes price directionality when observed price differences are not statistically significant. The third model introduces a more systematic framework, relaxing the seizure constraints at the state level, and the penalties for price relaxations are linked to the level of statistical evidence. Across the formulations, rather than identifying a single network, we generate a series of plausible trafficking scenarios, revealing how the inferred network changes under different parameter settings. All three optimization models were implemented in R and solved using the Gurobi Optimizer. Each model was solved in less than a second on a PC with an Intel i7-13700 processor (2.1 GHz) and 32 GB RAM.

4.3.1 Model 1: Relaxed Seizure Constraint

Using the above assumption and notation, we formulate the network discovery process as a mixed-integer programming (MIP) problem presented in Equation 4.1. Standard supervised learning frameworks minimize a loss function based on minimizing the differences between predicted values and actual outcomes. However, for illicit networks, there is no known ground truth. This means that

multiple network structures could be equally consistent with all or part of the available data. To facilitate network inference, we define an objective function that produces networks that are both simple and reflective of the illicit drug flows literature. Specifically, the objective function (4.1a) balances the two terms: the first counts the number of states with positive external inflow, and the second measures the total inter-state cocaine flow. We minimize the objective function, using a tuning parameter $\alpha \in [0, 1]$ that controls the trade-off between these two network characteristics.

$$\underset{s_i, t_i, w_{ij}}{\text{minimize}} \quad z = \alpha \sum_{i=1}^{N} t_i + (1-\alpha) \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \tag{4.1a}$$

$$\text{subject to} \quad \sum_{i=1}^{N} s_i = 100, \tag{4.1b}$$

$$s_i + \sum_{j=1}^{N} w_{ji} - \sum_{j=1}^{N} w_{ij} = 100 \cdot o_i / \sum_{j=1}^{N} o_j, \qquad i = 1, ..., N \tag{4.1c}$$

$$s_i + \sum_{j=1}^{N} w_{ji} \geq 100 \cdot \varepsilon \cdot c_i / \sum_{j=1}^{N} c_j, \quad i = 1, ..., N \tag{4.1d}$$

$$w_{ji} = 0, \qquad \forall i, j : p_i < p_j, i, j \in \{1, ..., N\} \tag{4.1e}$$

$$\frac{s_i}{\min_i 100 \cdot o_i / \sum_{j=1}^{N} o_j} - t_i \geq 0, \qquad i = 1, ..., N \tag{4.1f}$$

$$-0.01 \cdot s_i + t_i \geq 0, \qquad i = 1, ..., N \tag{4.1g}$$

$$s_i, w_{ij} \geq 0, \qquad i, j \in \{1, ..., N\} \tag{4.1h}$$

$$t_i \in \{0, 1\} \qquad i = 1, ..., N \tag{4.1i}$$

The constraints formalize the modeling assumptions introduced above. First, Constraint (4.1b)

scales the total cocaine inflow to 100 (without loss of generality), so that each state's inflow s_i can be interpreted as a percentage. Constraint (4.1c) requires that the net inflow (inflow minus outflow) for each state matches its scaled consumption, $100 \cdot o_i / \sum_{j=1}^{N} o_j$, which represents the state's percentage share of national overdose deaths. Constraint (4.1d) links total inflow to the state's maximum observed seizure, but because seizures are noisy indicators of actual flow, we relax this constraint using a parameter $\varepsilon \in [0, 1]$, requiring only that inflow meets or exceeds $\varepsilon \cdot c_i / \sum_{j=1}^{N} c_j$. Constraint (4.1e) limits the flow from state j to i to zero if the price of cocaine is lower in i compared to j, enforcing directional flow toward higher prices. Constraints (4.1f) and (4.1g) ensure that the binary variable t_i reflects whether a state receives any inflow ($s_i > 0$). Finally, s_i and w_{ij} are defined to be non-negative, and t_i is binary by construction.

4.3.1.1 Model 1 Results and Comparison with Anecdotal Evidence

To obtain feasible solutions of the optimization problem (4.1), the seizure constraint had to be significantly relaxed, setting ε to a maximum of 0.4. The balance parameter α was set to 0.9 to manage the trade-off between the number of source states and unrealistic long-distance flows. Figure 4.1 illustrates inferred cocaine trafficking flows with these parameter settings, where major inflows (and distribution across the U.S.) are seen from Florida (northward), Arizona (to California and eastward), and Texas. Smaller inflows enter through Michigan and some East Coast states.

Figure 4.2 presents cocaine routes that were reported in the 2011 National Drug Threat Assessment published by National Drug Intelligence Center (2011), or the DEA. Comparison of the inferred network with the DEA trafficking map reveals both overlaps and differences from the inferred network. While the DEA map highlights California as a key entry point, the model suggests Arizona was more significant

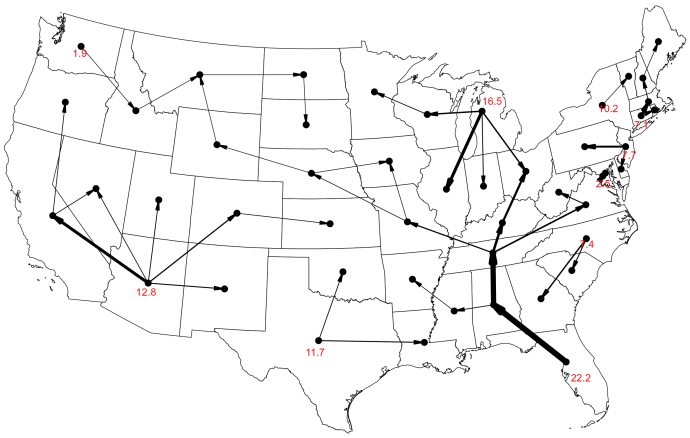


Figure 4.1: The optimal network from Model 1 with $\alpha = 0.9$ and $\varepsilon = 0.4$.

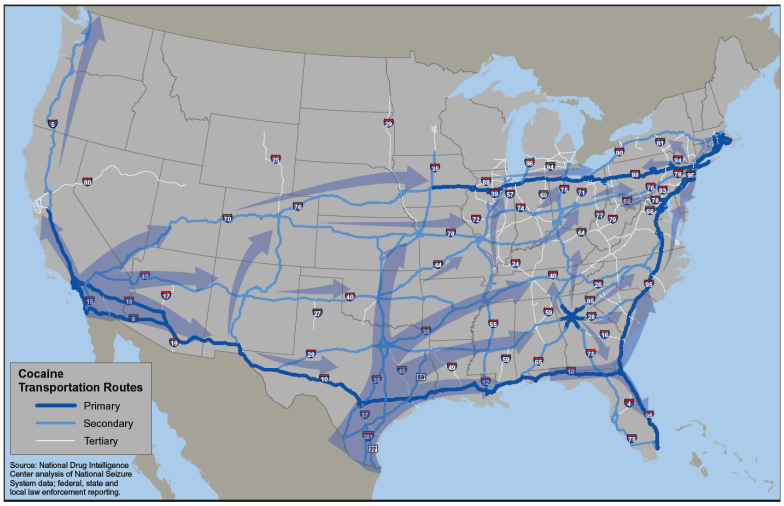


Figure 4.2: Routes for cocaine flows across the U.S.; source National Drug Intelligence Center (2011). Arrows indicate the direction of flows and their thickness represents flow magnitude. Source states are labeled with red numbers indicating the estimated amount of cocaine entering from outside the United States.

during the study period, possibly due to differences in border enforcement. The model also detects inflows from Michigan and eastern states that were not a part of the DEA map, highlighting how the data-driven method can provide additional or updated insights into trafficking patterns.

4.3.2 Model 2: Relaxed Seizure and Price Directionality

Model 1 (4.1) does not have a feasible solution without significantly relaxing the seizure constraint. In other words, a feasible solution satisfying the directionality of flow based on prices exists only when the required maximum seizures are reduced by 60%. In Model 1, we chose to relax the seizure constraint, as the seizure data are a noisy signal of the underlying network (maximum seizures are heavily influenced by law enforcement efforts) compared to price (median price estimates are based on multiple undercover purchases). However, if the prices in two adjacent states are statistically similar, we may want to relax the imposed directionality. More generally, it may be beneficial to trade off different signals to obtain different complementary views of the underlying network.

Therefore, in Model 2, we relax the price constraint when the level of confidence in the price directionality is low. More specifically, we remove the directionality constraints if the average purity-adjusted prices in two adjacent states are similar (assessed using Welch's t-test for equal means). If we fail to reject the null hypothesis at a selected significance level, we remove the directionality constraint between i and j, allowing cocaine flow in either direction between i and j but not both. The hypothesis testing reveals several state pairs (e.g., North Carolina and South Carolina, and Colorado and Utah) that have very similar purity-adjusted prices. Specifically, if statistical testing reveals that the price differences between i and j are not statistically significant, the constraint $w_{ij} = 0$ is removed from the model formulation.

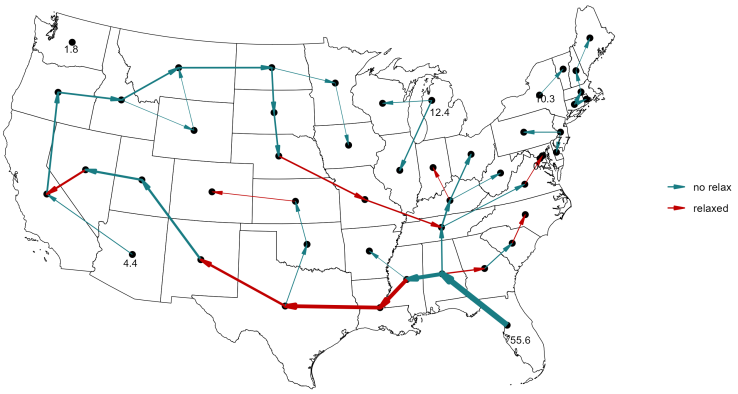


Figure 4.3: The optimal network from Model 2 with $\alpha = 0.9$ and $\varepsilon = 1.0$. Flows associated with the removed price constraints are colored red, while flows consistent with price directionality are colored green. The black numbers indicate the inflows.

As a demonstration, we rerun the analysis, keeping α at 0.9 and removing price constraints for state pairs with a t-test p-value above 0.25[2]. By allowing cocaine to flow in either direction between pairs of states with very similar prices, all the maximum seizure constraints (4.1d) are satisfied, and we therefore set ε equal to one. Figure 4.3 illustrates the resulting network.

Removing price directionality constraints that are not strongly supported by the price data results in different routes (compared to Model 1), reflecting alternative routes while satisfying the consumption and seizure information. The Model 2 network positions Alabama as a key supply state in the Southeast, with flows extending as far as North Carolina, which is no longer identified as a source state. The western portion of the network also changes: compared with Model 1, western flows become more closely connected to routes leading toward the eastern United States. On the West Coast, Washington has the same cocaine inflow but no longer sends cocaine to Idaho; instead, Idaho receives cocaine from Oregon.

[2]At this significance level, the constraints for 34% of neighboring state pairs (34 out of 101) are relaxed, reflecting the high level of noise in the data

4.3.3 Model 3: Statistical Price Penalty Relaxation

Model 2 relaxed the price directionality constraint using a fixed significance threshold applied to all neighboring state pairs. Under this approach, price directionality is either enforced or removed, depending on whether the observed price difference meets the selected threshold. However, the statistical evidence supporting price directionality varies across state pairs, which is not fully captured by a single binary decision.

To capture these differences, we update the formulation, resulting in Model 3, which assigns penalties to price directionality violations based on the strength of the statistical evidence. A violation receives a larger penalty when the data provide strong evidence that the price in state i exceeds the price in state j, and a smaller penalty when the observed price difference is small and therefore highly uncertain. This provides additional flexibility, discouraging violations of price directionality assumptions that are strongly supported by the data, but weakly supported price directionality results in a much smaller penalty. The goal of Model 3 is to provide a more flexible and statistically driven formulation by incorporating state-specific seizure relaxation and differentiated price directionality penalties based on statistical evidence.

To quantify the strength of the statistical evidence supporting price directionality from i to j, we consider the hypothesis test

$$H_0 : price_i \leq price_j, \quad H_1 : price_i > price_j.$$

Let $p_{p,ij}$ denote the corresponding p-value of the test. Small p-values imply stronger evidence that $price_i$ is greater than $price_j$, and therefore it is very unlikely for cocaine to flow from i to j,

whereas large p-values indicate weak evidence against such a flow. Consequently, violating the expected direction of flow should incur a larger penalty when $p_{p,ij}$ is small. Therefore, we set the penalty associated with the flow from i to j equal to $(1 - p_{p,ij})$ and incorporate it directly into the objective function.

Model 3 also incorporates state-specific seizure-relaxation variables, enabling the model to accommodate inconsistencies in seizure information across states. Let ε_i denote the seizure-relaxation variable for state i. A value of $\varepsilon_i = 1$ indicates that the maximum-seizure constraint is fully enforced, whereas $\varepsilon_i = 0$ corresponds to its maximum relaxation. The resulting formulation is given in Equation 4.2.

The objective function (4.2a) contains four components. The first sums the total amount of cocaine transported through the network, encouraging parsimonious networks with shorter and less extensive trafficking routes. The second term counts the number of source states, favoring solutions with fewer importation points. The third term accounts for the relaxation of the seizure constraints. Because larger values of ε_i correspond to less relaxation, the third term is subtracted from the objective function, thereby penalizing large relaxations and encouraging the optimal solution to satisfy as much of the observed seizure information as possible. Finally, the fourth term penalizes violations of the price directionality assumption. Specifically, a flow from state i to state j incurs a penalty of $(1 - p_{p,ij})$, where smaller p-values ($p_{p,ij}$) correspond to stronger statistical evidence that the price in state i exceeds that in state j. To avoid penalizing flows when the available data do not provide sufficient evidence for such a price difference, we set p-values above a prespecified threshold to one, before solving the optimization model (4.2). Consequently, only flows associated with sufficiently small p-values between two states i and j cost the objective function (4.2a), while flows with p-values greater than or equal to the threshold incur no relaxation cost.

The four terms are balanced through two nonnegative parameters, α and β, which can be set by the user to balance the number of importing states and overall flows.

$$\underset{s_i, t_i, w_{ij}, x_{ij}, z_{ij}, \varepsilon_i}{\text{minimize}} \quad z = \alpha \left(\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \right) + \beta \left(\sum_{i=1}^{N} t_i \right) - \sum_{i=1}^{N} \varepsilon_i + \sum_{i=1}^{N} \sum_{j=1}^{N} (1 - p_{p,ij})\, x_{ij} \tag{4.2a}$$

$$\text{subject to} \quad \sum_{i=1}^{N} s_i = 100, \tag{4.2b}$$

$$s_i + \sum_{j=1}^{N} w_{ji} - \sum_{j=1}^{N} w_{ij} = 100 \cdot o_i / \sum_{j=1}^{N} o_j, \qquad i = 1, ..., N \tag{4.2c}$$

$$s_i + \sum_{j=1}^{N} w_{ji} \geq 100 \cdot \varepsilon_i \cdot c_i / \sum_{j=1}^{N} c_j, \quad i = 1, ..., N \tag{4.2d}$$

$$w_{ij} \leq 100 x_{ij}, \qquad i, j \in \{1, ..., N\} \tag{4.2e}$$

$$w_{ij} \geq 0.0001 x_{ij}, \qquad i, j \in \{1, ..., N\} \tag{4.2f}$$

$$w_{ij} \leq 100 z_{ij}, \qquad i, j \in \{1, ..., N\} \tag{4.2g}$$

$$w_{ji} \leq 100(1 - z_{ij}), \qquad i, j \in \{1, ..., N\} \tag{4.2h}$$

$$s_i \leq 100 t_i, \qquad i = 1, ..., N \tag{4.2i}$$

$$s_i \geq 0.0001 t_i, \qquad i = 1, ..., N \tag{4.2j}$$

$$s_i, w_{ij} \geq 0, \qquad i, j \in \{1, ..., N\} \tag{4.2k}$$

$$t_i \in \{0, 1\} \qquad i = 1, ..., N \tag{4.2l}$$

$$x_{ij}, z_{ij} \in \{0, 1\} \qquad i, j \in \{1, ..., N\} \tag{4.2m}$$

$$\varepsilon_i \in [0, 1], \qquad i = 1, ..., N \tag{4.2n}$$

$$\alpha, \beta \in \mathbb{R}^+, \tag{4.2o}$$

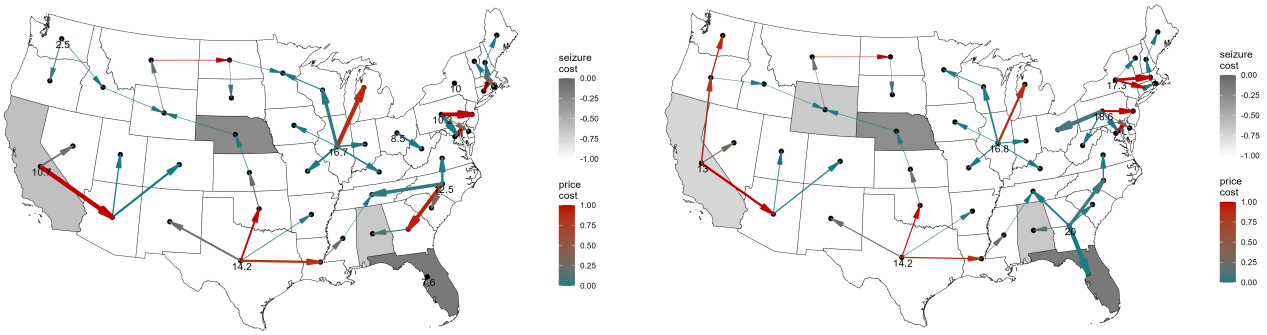
Constraint (4.2b) normalizes the total cocaine inflow entering the United States to 100 units. As in constraint (4.1b), this normalization allows all flows to be interpreted as percentages of the total supply. Constraint (4.2c) enforces flow conservation at each state. Constraint (4.2d) requires the total cocaine entering a state to be at least as large as its normalized maximum observed seizure. Unlike the previous model (4.1), no global seizure-relaxation parameter is used; instead ε_i is used to relax the seizure requirement for a state i.

Constraint (4.2e) and (4.2f) link the binary variables x_{ij} to the existence of a flow between states i and j of magnitude w_{ij}. Specifically, $x_{ij} = 1$ if and only if a positive flow exists from state i to state j, and $x_{ij} = 0$ otherwise. The variable x_{ij} is included in the objective function to ensure that penalties associated with relaxing the price directionality assumption are incurred only when a flow actually occurs between the two states. Therefore, state pairs with no flow ($w_{ij} = 0$) do not contribute to the price-relaxation penalty.

Constraints (4.2g) and (4.2h) ensure that at most one directional flow exists between any pair of neighboring states. Constraints (4.2i) and (4.2j) ensure that the source-state indicator t_i equals 1 if and only if state i imports cocaine from outside of the United States, and is zero otherwise. Finally, the remaining constraints specify non-negativity and integrality requirements for the flow and indicator variables.

4.3.3.1 Model 3 Results

We present two optimal solutions generated using $\alpha = 1$, while varying the parameter β. Figure 4.4(a) shows the solution obtained with $\beta = 5$, whereas Figure 4.4(b) presents the solution obtained with a larger weight ($\beta = 10$) on the number of importing states. In both figures, the extent of



(a) With $\alpha = 1, \beta = 5$

(b) With $\alpha = 1, \beta = 10$

Figure 4.4: The optimal network for Model 3 with two different parameter settings (a) $\alpha = 1$ and $\beta = 5$, (b) $\alpha = 1$ and $\beta = 10$. State-level seizure relaxation (ε_i) is indicated by gray shading. Price directionality violation penalties are represented by flow colors ranging from green to red, with redder flows indicating higher penalty for relaxing price directionality.

seizure and price constraint relaxation is represented by the corresponding penalty costs.

The two networks exhibit similar patterns in domestic cocaine flows. Especially, near Illinois and Texas, both states supply cocaine to several neighboring states. However, when $\beta = 10$, the model suggests a northward flow from California to Washington, which is not present when $\beta = 5$. This occurs because, when $\beta = 10$, the number of importing states is penalized more and Washington is no longer selected as an importing state. The southeastern network also changes, primarily because the regional source shifts from North Carolina to Georgia.

Overall, both Model 3 solutions exhibit noticeably different trafficking patterns compared to the solutions of Models 1 and 2 (Figures 4.1 and 4.3). In particular, in the Model 3 solutions, the cocaine flows originating from California become shorter despite a similar number of importing states. Further, Florida no longer serves as a key source state of cocaine, and under the larger value $\beta = 10$, is not selected as an importing state, as shown in Figure 4.4(b).

Another notable change occurs in the Midwest, where in the Model 3 solutions, Michigan is no longer identified as a source state, and Illinois emerges as a regional supplier. At the same time,

California, Texas, and New York remain major importing states for both parameter settings, suggesting that their roles are robust to the introduction of the relaxation mechanism.

The Model 3 results (Figure 4.4) demonstrate that incorporating state-level seizure constraints and statistical penalties for violating the price directionality assumption produces different plausible cocaine supply chain networks. Rather than forcing all flows to strictly follow observed price gradients, the model relaxes directionality assumptions by applying penalties proportional to the strength of the statistical evidence (rather than the binary relaxation used in Model 2), resulting in alternative trafficking routes. This flexibility is particularly valuable in settings such as ours, where the available data exhibit substantial variability and uncertainty.

Together, these two relaxation mechanisms reduce the need to fit potentially noisy observations exactly in the optimization model. Instead, Model 3 prioritizes network structures that are as consistent as possible with the available evidence, while penalizing differences based on statistical principles. The persistence of several major importing states across parameter settings further suggests that certain structural features of the inferred network are relatively robust, whereas the locations of intermediate suppliers and specific trafficking routes remain sensitive to assumptions regarding price directionality and seizure enforcement patterns. These findings highlight the importance of incorporating uncertainty into network inference models and demonstrate how carefully designed relaxation mechanisms can improve the plausibility and provide an alternative lens of inferred illicit supply chains.

4.4 Conclusion

This study developed a data-driven optimization framework for inferring hidden cocaine trafficking networks in the United States using multiple incomplete and noisy data sources. Unlike traditional

network inference approaches that require at least partial knowledge of the underlying network structure, the proposed method operates without any observed network links. Instead, we leverage information from cocaine prices, seizures, and overdose deaths to identify plausible trafficking routes. By formulating the problem as a network optimization model, the approach integrates multiple sources of evidence while explicitly accounting for the uncertainty and inconsistency that are inherent in illicit drug market data.

The empirical results demonstrate that the proposed framework can generate network structures that are broadly consistent with available anecdotal evidence while also revealing alternative trafficking patterns supported by the data. The inferred networks consistently identify several major importing states and highlight the importance of regional redistribution hubs within the domestic cocaine supply chain. Furthermore, the comparison between alternative formulations illustrates how different assumptions regarding price directionality and seizure information can lead to different but plausible network configurations. Rather than producing a single definitive network, the different parameters and formulations generate a set of evidence-based scenarios that can help law enforcement agencies and policymakers evaluate alternative hypotheses about trafficking activity.

Several limitations remain. First, cocaine prices may be influenced by factors other than transportation costs and trafficking risk. For example, differences in local demand, competition from substitute drugs, market structure, or enforcement intensity may cause cocaine prices to decrease along a trafficking route. In a state where demand for cocaine is relatively low, cocaine transported from another state may be sold at a lower price than in the supplying state. The current price-directionality assumption does not capture these patterns. Although the formulations allow price directionality to be relaxed, they do not explicitly model the underlying economic factors that produce these exceptions.

Second, our model does not use cocaine purity as an additional source of directional information.

Because cocaine purity generally decreases as the drug moves through the supply chain and is diluted before reaching retail markets, purity information could provide additional constraints on the likely direction of cocaine flows. Incorporating such information may improve network identification and further reduce ambiguity in the inferred routes. However, adding additional constraints inevitably reduces the feasible region of the optimization problem and may increase the difficulty of obtaining feasible and interpretable solutions when the underlying data are noisy.

Future research should therefore focus on developing more effective methods for balancing multiple sources of information while preserving model feasibility, interpretability, and consistency with data. This includes incorporating additional constraints, exploring alternative formulations for handling uncertainty, and developing systematic procedures for selecting relaxation parameters and penalty weights. More broadly, the proposed framework could be extended to other illicit supply chains and hidden networks where direct observations of network structure are unavailable. By combining optimization, statistical evidence, and multiple noisy data sources, such approaches have the potential to provide valuable insights into a wide range of clandestine systems.

Chapter 5: Conclusion

Illicit supply chains present a unique challenge for researchers and policymakers because the underlying networks are intentionally concealed. Unlike legitimate supply chains, where transactions, inventories, and transportation flows are routinely recorded, illicit activities generate only incomplete and partial observations. As a result, analysts must often redesign conventional methodologies and draw inferences to incorporate sparse, incomplete, and inconsistent data while operating with little or no ground-truth information. These limitations complicate efforts to understand trafficking behavior and develop effective interdiction strategies.

This dissertation addressed these challenges through three complementary studies focused on cocaine trafficking in Colombia and the United States. Although each study examined a different analytical problem, all three shared a common objective: improving the extraction of actionable information from imperfect data on illicit supply chains.

The first study investigated regional drivers of cocaine trafficking in Colombia using partially observed information on trafficking routes. To account for substantial spatial heterogeneity and localized data sparsity, a geographically weighted logistic regression framework incorporating Firth's penalized maximum likelihood estimation was developed. The results demonstrated that the relationships between trafficking destinations and associated factors vary considerably across regions.

The second study considered situations in which no direct information on trafficking routes is available and investigators must rely on secondary indicators such as drug seizures. Using crack cocaine seizure data in the United States, the study identified an important limitation of conventional LISA analysis when applied to zero-heavy criminal datasets. To address this issue, an improved LISA method was developed to enhance the identification and interpretation of spatial clusters with zero-heavy data.

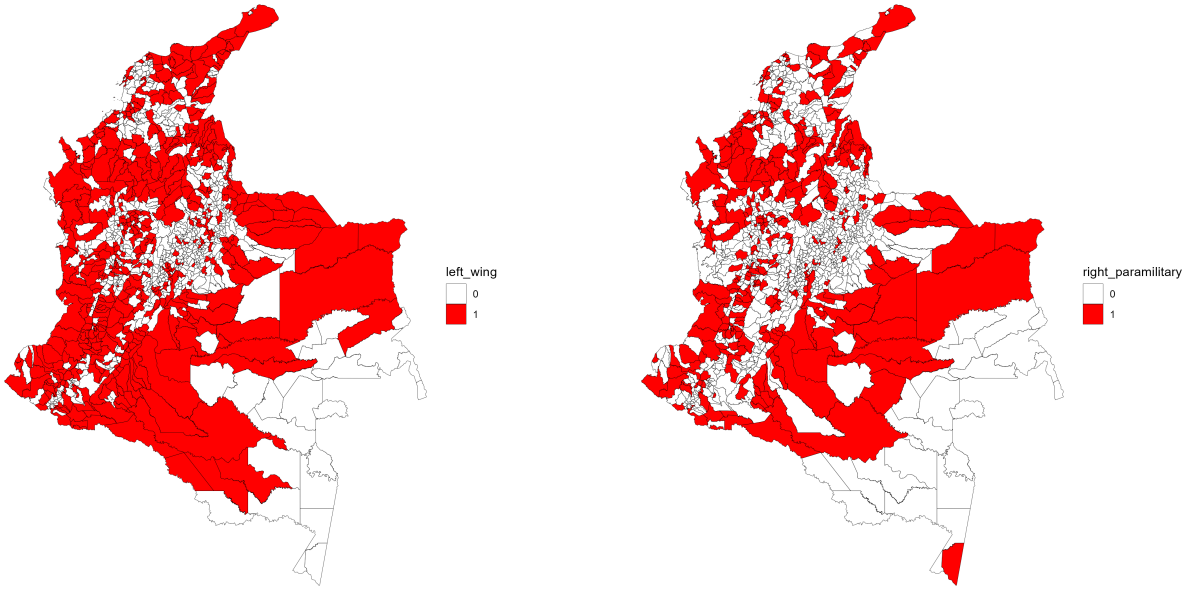
The results demonstrated that conventional methods may overlook meaningful spatial patterns and that appropriate adjustments can substantially improve hot spot detection.

The third study addressed the problem of inferring cocaine trafficking networks from multiple incomplete and potentially inconsistent variables. A novel data-driven optimization framework was developed to infer plausible cocaine supply-chain structures in the United States, supported by multiple incomplete and potentially inconsistent data sources, including seizure records, undercover purchase data, and overdose mortality. Rather than producing a single deterministic network, the framework generates multiple candidate network structures that are consistent with observed evidence, thereby providing decision-makers with a richer understanding of uncertainty in network inference. The resulting approach demonstrates how optimization methods can be used to reconstruct hidden illicit supply chains when direct observations are unavailable.

Overall, this dissertation demonstrates that meaningful insights into illicit drug supply chains can be obtained without complete information. Across all three studies, the central methodological theme is the development of analytical approaches that explicitly account for sparsity, incompleteness, and uncertainty. Collectively, these studies show that these characteristics should not be viewed merely as limitations of criminal data but rather as features that must be incorporated directly into model design and interpretation to fully utilize the data and support more effective enforcement decisions.

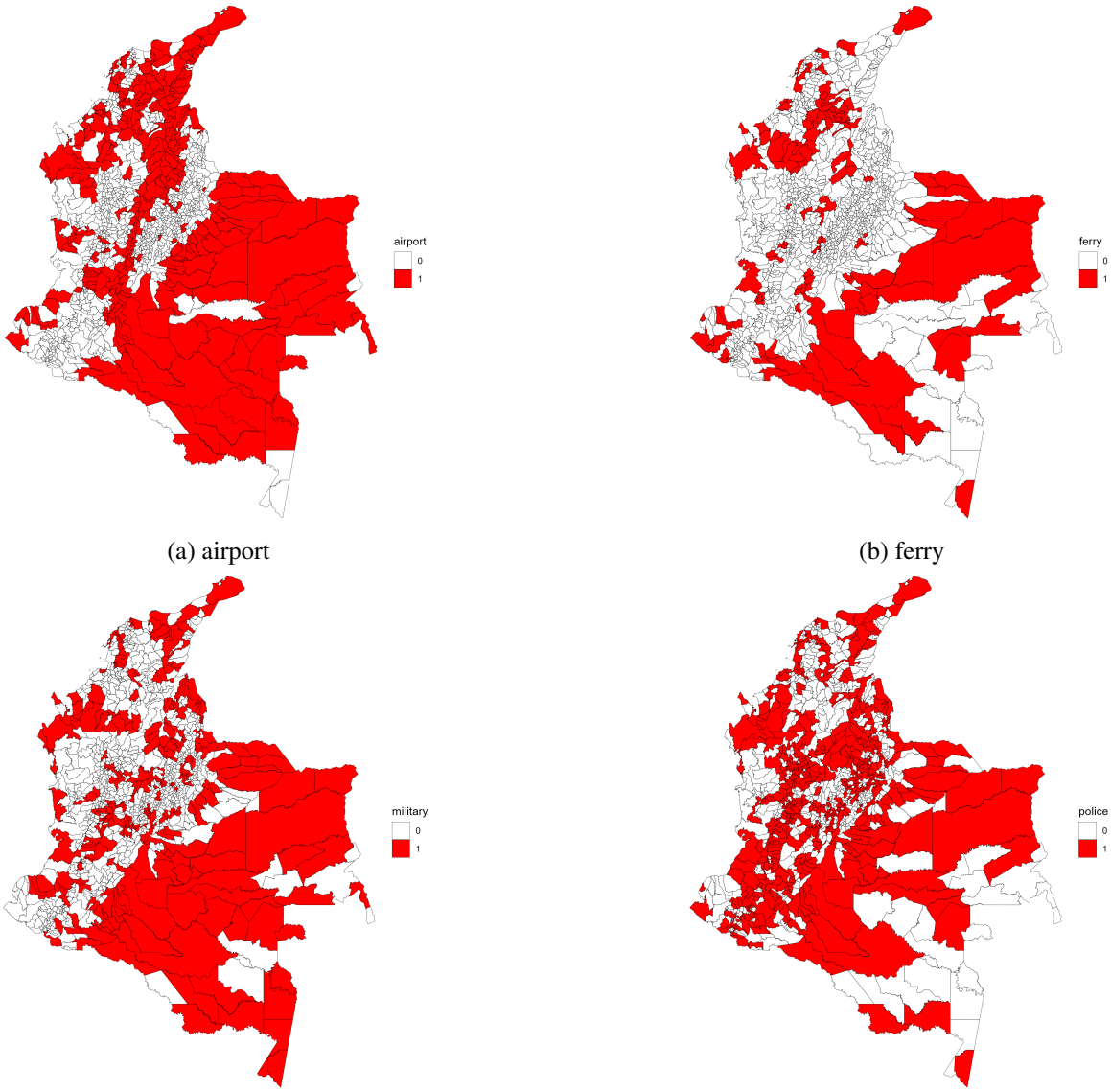
Appendix A: Geographical Variation of Aggregated Data: Remaining Variables

This appendix presents maps of the independent variables used in the second study that were not included in Section 2.2.3.



(a) left_wing (b) right_paramilitary

Figure A.1: Geographical variation of (a) left_wing and (b) right_paramilitary data.



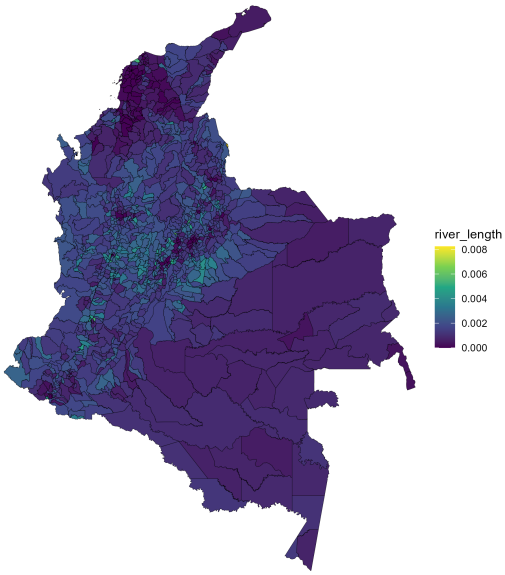
(a) airport

(b) ferry

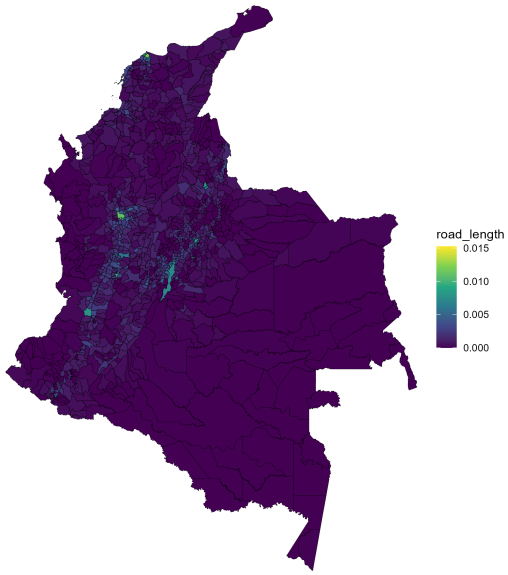
(c) military

(d) police

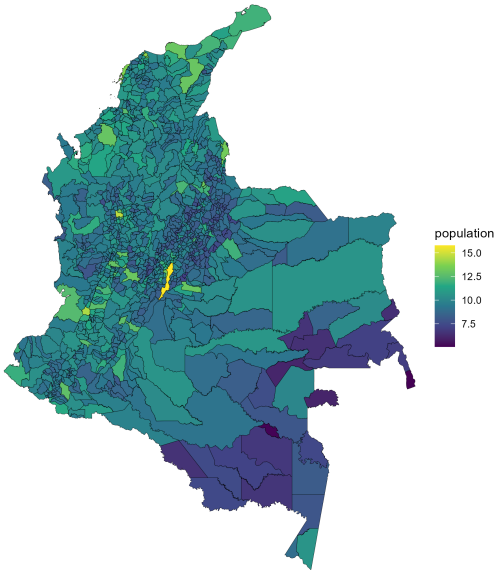
Figure A.2: Geographical variation of (a) airport, (b) ferry, (c) military, and (d) police data.



(a) river_length



(b) road_length



(c) population (log values)

Figure A.3: Geographical variation of (a) river_length, (b) road_length, and (c) population data.

Appendix B: Firth's Penalized Likelihood Method

When estimation is based on a relatively small sample, models may suffer from multicollinearity and unstable coefficient estimates (Wheeler and Tiefelsdorf 2005), including local GWR models. To address these issues, several extensions have incorporated regularization techniques, including ridge- and LASSO-based geographically weighted models (Bárcena et al. 2014, Comber and Harris 2018, Wang and Li 2017, Wheeler 2007, 2009, Yoneoka et al. 2016). However, these approaches do not explicitly consider (quasi-)complete separation when applying geographically weighted logistic regression, which may lead to unstable or inflated coefficient estimates (Heinze and Schemper 2002).

To address the severe class imbalance in the data, we propose a GWR framework that incorporates Firth's penalized maximum likelihood (PML) estimator, a bias-reduction technique that has been widely applied in conventional logistic regression settings (Heinze 2006, Heinze et al. 2013, Rainey 2016). Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_M)$ for $M \in \mathbb{N}$ denote the vector of model parameters and let $L(\boldsymbol{\theta})$ denote the likelihood function. Standard maximum likelihood estimation (MLE) seeks the parameter values that maximize the likelihood, or equivalently, the log-likelihood $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$. The gradient of the log-likelihood, $U(\boldsymbol{\theta}) \equiv \nabla\ell(\boldsymbol{\theta})$, is referred to as the *score vector* and measures the sensitivity of the log-likelihood to changes in the model parameters. The standard maximum likelihood estimators are obtained by solving the *score equation*

$$U(\boldsymbol{\theta}) = \nabla\ell(\boldsymbol{\theta}) = \mathbf{0}. \tag{B.1}$$

By adjusting the standard score equation to account for estimator bias, Firth's PML method produces finite and more stable coefficient estimates, particularly in the presence of complete or quasi-

complete separation, where conventional maximum likelihood estimation may fail or yield excessively large coefficients (Heinze 2006, Heinze and Schemper 2002).

Firth's method modifies the standard MLE by introducing a bias-reducing penalty term into the score function for M parameters $\boldsymbol{\theta}$:

$$U^*(\boldsymbol{\theta}) = U(\boldsymbol{\theta}) - I(\boldsymbol{\theta})b(\boldsymbol{\theta}) = 0 \quad \text{(Firth's score)} \tag{B.2}$$

where $I(\boldsymbol{\theta})$ is the Fisher information matrix (the variance of the standard score (B.1):

$$[I(\boldsymbol{\theta})]_{j,k} = E\left[\left(\frac{\partial}{\partial\theta_j}\ell(\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\theta_k}\ell(\boldsymbol{\theta})\right)\middle|\,\boldsymbol{\theta}\right] \tag{B.3}$$

for $j, k \in \{1, 2, ..., M\}$), and $b(\boldsymbol{\theta})$ represents the biases for the estimators of $\boldsymbol{\theta}$. This penalized maximum likelihood approach yields more stable coefficient estimates compared to standard MLE, particularly in the presence of separation in datasets (Heinze 2006, Heinze and Schemper 2002).

For a logistic regression, consider the following model with a dependent variable y and M independent variables $\mathbf{x} = (x_1 x_2 \cdots x_M)$ and $x_0 = \mathbf{1}$:

$$P(y|\mathbf{x}, \boldsymbol{\theta}) = \pi = \left[1 + \exp\left(-\sum_{m=0}^{M} x_m\theta_m\right)\right]^{-1}. \quad \text{(logistic regression model)} \tag{B.4}$$

The standard score function (B.1) corresponding to model (B.4) for a parameter θ_m is:

$$U(\theta_m) = \sum_{i=1}^{n}(y_i - \pi_i)x_{mi} = 0 \tag{B.5}$$

For logistic regression, Firth's score (B.3) for a parameter θ_m becomes:

$$U^*(\theta_m) = \sum_{i=1}^{n} \left[(y_i - \pi_i) + h_i \left(\frac{1}{2} - \pi_i \right) \right] x_{mi} = 0, \tag{B.6}$$

where h_i denotes the ith diagonal element of the hat matrix $H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2}$ and $W = \text{diag}(\pi_i(1 - \pi_i)$. Equation (B.6) can be rewritten as follows:

$$U^*(\theta_m) = \sum_{i=1}^{n} \left[(y_i - \pi_i) \left(1 + \frac{h_i}{2} \right) + ((1 - y_i) - \pi_i) \frac{h_i}{2} \right] x_{mi} = 0 \tag{B.7}$$

This representation shows that the adjusted score (B.7) contains contributions from both the observed response y_i and its binary complement $(1 - y_i)$ for the same observation i (Heinze and Schemper 2002). As noted by Heinze and Schemper (2002), this formulation can be interpreted as augmenting the data with pseudo-observations, thereby preventing coefficient estimates from diverging under complete or quasi-complete separation and "guarantees finite estimate" for θ_m.

Bibliography

ACAPS (2021). Colombia - regional needs analysis: Chocó. Technical report, Assessment Capacities Project.

Ahmad, F., C. JA, R. LM, and S. P (2024). Provisional drug overdose death counts. Technical report, National Center for Health Statistics.

Anderson, D., G. Benitez, M. Bjarnadottir, and S. Reyya (2025). Using LLMs to create analytical datasets: A case study of reconstructing the historical memory of Colombia. *arXiv preprint arXiv:2509.04523*.

Andrade, F. H. E., C. J. Molina, and R. G. Llave (2024). Analysis of illicit drug trafficking in Colombia's maritime spaces: A spatial exploratory approach. *Journal of Maritime Research 21*(2), 135–145.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis 27*, 93–115.

Anselin, L. (2019). A local indicator of multivariate spatial association: extending Geary's c. *Geographical Analysis 51*(2), 133–150.

Arkes, J., R. L. Pacula, S. M. Paddock, J. P. Caulkins, and P. Reuter (2008). Why the DEA STRIDE Data are Still Useful for Understanding Drug Markets. NBER Working Papers 14224, National Bureau of Economic Research, Inc.

Atkinson, P. M., S. E. German, D. A. Sear, and M. J. Clark (2003). Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression. *Geographical Analysis 35*(1), 58–82.

Bárcena, M. J., P. Menéndez, M. B. Palacios, and F. Tusell (2014). Alleviating the effect of collinearity in geographically weighted regression. *Journal of Geographical Systems 16*(4), 441–466.

Benítez, G. J., S. Chandra, T. C. L. W. Cuadros Veloza, and I. J. D. Díaz Cárdenas (2019). Following the price: Identifying cocaine trafficking networks in Colombia. *Global Crime 20*(2), 90–114.

Bjarnadottir, M. V., S. Chandra, P. He, and G. Midgette (2024). Analyzing illegal psychostimulant trafficking networks using noisy and sparse data. *IISE Transactions 56*(3), 269–281.

Brunsdon, C., A. S. Fotheringham, and M. E. Charlton (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis 28*(4), 281–298.

Cai, X., Z. He, and C. Wen (2025). Does urban renewal program increase urban vitality? causal evidence from Beijing City, China. *Applied Geography 183*, 103732.

Caulkins, J. P. and P. Reuter (1998). What price data tell us about drug markets. *Journal of Drug Issues 28*(3), 593–612.

Chandra, S., M. Liu, M. V. Bjarnadottir, and G. Midgette (2025). A novel method for measuring the extent of cocaine flows across the US. *Justice Quarterly 42*(7), 1518–1541.

Clauset, A., C. Moore, and M. E. Newman (2008). Hierarchical structure and the prediction of missing links in networks. *Nature 453*(7191), 98–101.

Cohen, J. and G. Tita (1999). Diffusion in homicide: Exploring a general method for detecting spatial diffusion processes. *Journal of Quantitative Criminology 15*, 451–493.

Comber, A. and P. Harris (2018). Geographically weighted elastic net logistic regression. *Journal of Geographical Systems 20*(4), 317–341.

Comber, A., Y. Wang, Y. Lü, X. Zhang, and P. Harris (2018). Hyper-local geographically weighted regression: extending GWR through local model selection and local bandwidth optimization. *Journal of Spatial Information Science* (17), 63–84.

D'Agostino Jr, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in medicine 17*(19), 2265–2281.

DANE (2018). Departamento Administrativo Nacional de Estadística: Censo Nacional de Población y Vivienda 2018. https://www.dane.gov.co/files/censo2018/informacion-tecnica/PERSONAS_DEMOGRAFICO_Cuadros_CNPV_2018.xlsx. Accessed: 2026-06-05.

Dávalos, E. and L. F. Morales (2023). Diffusion of crime control benefits: Forced eradication and coca crops in Colombia. *Spatial Economic Analysis 18*(3), 292–317.

Davis, M. L., B. Neelon, P. J. Nietert, L. F. Burgette, K. J. Hunt, A. B. Lawson, and L. E. Egede (2021). Propensity score matching for multilevel spatial data: accounting for geographic confounding in health disparity studies. *International Journal of Health Geographics 20*(1), 10.

Doreian, P., V. Batagelj, and A. Ferligoj (2005). *Generalized blockmodeling*. Cambridge, UK: Cambridge University Press.

Drug Enforcement Administration (2024). National drug threat assessment 2024. Technical report, U.S. Department of Justice.

Drug Enforcement Administration, U.S. Department of Justice (2020). *National Drug Threat Assessment*.

Du, X., B. Hansen, S. Zhang, and E. Zou (2026). Coca's return and the American overdose fallout. Technical report, National Bureau of Economic Research.

Feng, C. X. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of statistical distributions and applications 8*(1), 8.

Feng, C. X. and C. B. Dean (2012). Joint analysis of multivariate spatial count and zero-heavy count outcomes using common spatial factor models. *Environmetrics 23*(6), 493–508.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*(1), 27–38.

Fotheringham, A. S., C. Brunsdon, and M. Charlton (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. New York, NY: Wiley.

Fotheringham, A. S., W. Yang, and W. Kang (2017). Multiscale geographically weighted regression (MGWR). *Annals of the American Association of Geographers 107*(6), 1247–1265.

Friston, K. J., B. Li, J. Daunizeau, and K. E. Stephan (2011). Network discovery with DCM. *Neuroimage 56*(3), 1202–1221.

Getis, A. and J. K. Ord (1992). The analysis of spatial association by use of distance statistics. *Geographical analysis 24*(3), 189–206.

González, H. Z. (2022). *Coca, cocaína y violencia*, Chapter 5. Universidad de los Andes.

Grajales-Marín, A. F., F. H. Sepúlveda-Murillo, A. Tapia, and A. Tabares (2025). Spatial analysis of socioeconomic data and its relationship with illicit crops in Nariño-Colombia. *PloS one 20*(1), e0316709.

Gutiérrez Sanín, F. (2008). Telling the difference: Guerrillas and paramilitaries in the colombian war. *Politics & Society 36*(1), 3–34.

Haining, R., J. Law, and D. Griffith (2009). Modelling small area counts in the presence of overdispersion and spatial autocorrelation. *Computational Statistics and Data Analysis 53*(8), 2923–2937.

He, Y., Y. Zhao, and K. L. Tsui (2021). An adapted geographically weighted LASSO (Ada-GWL) model for predicting subway ridership. *Transportation 48*(3), 1185–1216.

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in medicine 25*(24), 4216–4226.

Heinze, G., M. Ploner, and J. Beyea (2013). Confidence intervals after multiple imputation: combining profile likelihood information from logistic regressions. *Statistics in medicine 32*(29), 5062–5076.

Heinze, G. and M. Schemper (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine 21*(16), 2409–2419.

Holme, P. and M. Huss (2005). Role-similarity based functional prediction in networked systems: application to the yeast proteome. *Journal of the Royal Society Interface 2*(4), 327–333.

Holmes, J. S., S. A. G. De Pineres, and K. M. Curtin (2006). Drugs, violence, and development in Colombia: A department-level analysis. *Latin American Politics and Society 48*(3), 157–184.

Horowitz, J. (2001, December). Should the DEAs STRIDE Data Be Used for Economic Analyses of Markets for Illegal Drugs? *Journal of the American Statistical Association 96*, 1254–1271.

Hughes, R. and D. MacKenzie (2016). Transportation network company wait times in greater Seattle, and relationship to socioeconomic indicators. *Journal of Transport Geography 56*, 36–44.

Ibanez, M. and P. Martinsson (2013). Curbing coca cultivation in Colombia—a framed field experiment. *Journal of Public Economics 105*, 1–10.

Indepaz (2008-2020). Report on presence of narcoparamilitary groups. Technical report, Institute of Studies for Development and Peace of Colombia.

Jiménez-García, W. G., W. Arenas-Valencia, and N. Bohorquez-Bedoya (2023). Violent drug markets: Relation between homicide, drug trafficking and socioeconomic disadvantages: A test of contingent causation in Pereira, Colombia. *Social Sciences 12*(2), 54.

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research 23*(1), 69–86.

Kenney, M. (2007). The architecture of drug trafficking: Network forms of organisation in the Colombian cocaine trade. *Global Crime 8*(3), 233–259.

Kolenikov, S. (2011). Biases of parameter estimates in misspecified structural equation models. *Sociological Methodology 41*(1), 119–157.

LaFree, G., L. Dugan, M. Xie, and P. Singh (2012). Spatial and temporal patterns of terrorist attacks by ETA 1970 to 2007. *Journal of Quantitative Criminology 28*(1), 7–29.

Li, T., I. Perrigne, and Q. Vuong (2002). Structural estimation of the affiliated private value auction model. *The RAND Journal of Economics 33*(2), 171–193.

Liu, Y., D. Tong, and X. Liu (2015). Measuring spatial autocorrelation of vectors. *Geographical Analysis 47*(3), 300–319.

Lü, L. and T. Zhou (2011). Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications 390*(6), 1150–1170.

Mateus-Rugeles, A. and P. C. Arias (2020). Narcotraffic as connected political crime in colombia. *The University of Miami Inter-American Law Review 51*(2), 1–62.

Mejia, D. and P. Restrepo (2016). The economics of the war on illegal drug production and trafficking. *Journal of Economic Behavior & Organization 126*, 255–275.

Milenioscuro (2014). Mapa de colombia (regiones naturales). Wikimedia Commons. Licensed under CC BY-SA 4.0; Accessed: 2026-07-29.

Millán-Quijano, J. (2020). Internal cocaine trafficking and armed violence in Colombia. *Economic Inquiry 58*(2), 624–641.

Musalem, A., M. Olivares, E. T. Bradlow, C. Terwiesch, and D. Corsten (2010). Structural estimation of the effect of out-of-stocks. *Management Science 56*(7), 1180–1197.

National Drug Intelligence Center (2011). National drug threat assessment 2011. Technical report, Drug Enforcement Administration, U.S. Department of Justice.

National Drug Intelligence Center (2016). National drug threat assessment 2016. Technical report, Drug Enforcement Administration, U.S. Department of Justice.

Negret, P. J., L. Sonter, J. E. Watson, H. P. Possingham, K. R. Jones, C. Suarez, J. M. Ochoa-Quintero, and M. Maron (2019). Emerging evidence that armed conflict and coca cultivation influence deforestation patterns. *Biological Conservation 239*, 108176.

Norman, S. V. (2018). Narcotization as security dilemma: The farc and drug trade in colombia. *Studies in Conflict & Terrorism 41*(8), 638–659.

ODC (1997-2022). Drug Observatory of Colombia. https://www.minjusticia.gov.co/programas/observatorio-de-drogas-de-colombia. Accessed: 2026-06-05.

Oldekop, J. A., K. R. E. Sims, and B. K. Karn (2019). Reductions in deforestation and poverty from decentralized forest management in Nepal. *Nature Sustainability 2*, 421–428.

OpenStreetMap (2026). Openstreetmap data: Colombia. `https://download.geofabrik.de/south-america/colombia.html`. Accessed: 2026-06-05.

Rainey, C. (2016). Dealing with separation in logistic regression models. *Political Analysis 24*(3), 339–355.

Reyes, L. C. (2014). Estimating the causal effect of forced eradication on coca cultivation in Colombian municipalities. *World Development 61*, 70–84.

Rivera, R. L. (2013). *Delimitación de las regiones naturales de Colombia (Natural regions of Colombia)*. Sociedad Geográfica de Colombia, Academia de Ciencias Geográficas.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rouse, S. M. and M. Arce (2006). The drug-laden balloon: Us military assistance and coca production in the central andes. *Social Science Quarterly 87*(3), 540–557.

Ruiz, A. R., H. L. Correa, D. O. León, and S. Williams (2016). Coca cultivation and crop eradication in Colombia: The challenges of integrating rural reality into effective anti-drug policy. *International Journal of Drug Policy 33*, 56–65.

Ruiz, A. R., U. Pascual, and M. Romero (2013). An exploratory spatial analysis of illegal coca cultivation in Colombia using local indicators of spatial association and socioecological variables. *Ecological indicators 34*, 103–112.

Sierra, J. R. (2016). La narcotización del activismo guerrillero de las farc y el eln 1998-2012. *Revista Unisci* (41), 205–233.

Strebulaev, I. A. and T. M. Whited (2012). Dynamic models and structural estimation in corporate finance. *Foundations and Trends in Finance 6*(1–2), 1–163.

Tao, R. and Y. Chen (2023). Applying local indicators of spatial association to analyze longitudinal data: the absolute perspective. *Geographical Analysis 55*(2), 225-238.

UNODC (1999-2023). United nations office on drugs and crime illicit crop monitoring.

UNODC (2021). *Cocaine Insights 2*. Vienna.

UNODC (2024). *World drug report*. United Nations publication.

UNODC (2026). *Research Brief on Transnational Organized Crime*. United Nations publication.

Wang, Q., J. Ni, and J. Tenhunen (2005). Application of a geographically-weighted regression analysis to estimate net primary production of Chinese forest ecosystems. *Global ecology and biogeography 14*(4), 379–393.

Wang, W. and D. Li (2017). Structure identification and variable selection in geographically weighted regression models. *Journal of Statistical Computation and Simulation 87*(10), 2050–2068.

Washington/Baltimore High Intensity Drug Trafficking Area (2022). High Intensity Drug Trafficking Area Annual Report.

Wheeler, D. and M. Tiefelsdorf (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems 7*(2), 161–187.

Wheeler, D. C. (2007). Diagnostic tools and a remedial method for collinearity in geographically weighted regression. *Environment and planning A 39*(10), 2464–2481.

Wheeler, D. C. (2009). Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso. *Environment and planning A 41*(3), 722–742.

Yang, Y. and S. B. Green (2010). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling: A Multidisciplinary Journal 17*(1), 66–81.

Yoneoka, D., E. Saito, and S. Nakaoka (2016). New algorithm for constructing area-based index with geographical heterogeneities and variable selection: An application to gastric cancer screening. *Scientific reports 6*(1), 26582.