THE UNIVERSITY OF CHICAGO

SPATIAL HETEROGENEITY IN INTRA-STATE CONFLICTS

A THESIS SUBMITTED TO
THE FACULTY OF THE COMMITTEE ON QUANTITATIVE METHODS AND SOCIAL SCIENCE ANALYSIS (QMSA)

A paper submitted in partial fulfillment of the requirements
for the Master of Arts degree in the Master of Arts Program in the Social Sciences

BY
SERENA BERNSTEIN

Faculty Advisor: Paul Poast
Preceptor: Jingyuan Qian

CHICAGO, ILLINOIS
JUNE 2025

TABLE OF CONTENTS

CHAPTER 1

ABSTRACT

This paper addresses a key gap in the spatial analysis of sub-national conflict: the under utilization of spatial heterogeneity methods. Using data on the Kurdish conflict from 1980 to 2015, I demonstrate how standard Ordinary Least Squares (OLS) and spatial regression techniques often fail to uncover meaningful relationships due to unaccounted for spatial nonstationarity. By applying spatially constrained endogenous regime models, I identify distinct regions where conflict is significantly shaped by the presence of oil—the primary predictor in this case study. Further analysis reveals a significant spatial lag effect within one of these regions. These findings highlight the critical importance of incorporating spatial heterogeneity into conflict modeling and offer a practical methodological framework for doing so in future research.

CHAPTER 2

INTRODUCTION

Over the past two decades, geographic analysis has increasingly entered the lexicon of International Relations and the study of Intra-State Conflict. Thanks to the popularization of accessible GIS softwares and the proliferation of novel spatial datasets (ACLED, GeoEPR, and CShapes, to name just a few) [22, 25, 33], researchers are increasingly able to investigate the role of geography in shaping the dynamics of conflict.

As is most common in the field of International Relations, the majority of geographic analysis is at the country level, either encompassing the whole state or aggregating along specific subnational administrative boundaries. These analyses are fit to borders both modern and historic. Following increased interest from scholars in the study of ethnicity and conflict, the field expanded its geographic toolbox to include spatially defined ethnic groups as units of analysis [33]. Traditionally, examining geographic measures such as distance to a border/capital or percent mountainous terrain, these papers have realized significant findings, shedding light on which geographic factors meaningfully impact conflict propensity.

What is much less common has been the study of microgeography in conflict studies. Often implemented using a high-resolution spatial grid, these data were scarce and unreliable due to the measurement of small areas exogenous to the political boundaries from which data are normally collected. With a few notable exceptions [7, 24], remote sensing technology has only advanced more recently so that there are a sufficient number of usable datasets suitable for conducting these analyzes.

As has been pointed out by scholars like O'Loughlin and Raleigh [19, 24], grid-based data structures, when deployed with statistically appropriate methods, are incredibly powerful tools in the study of intra-state conflicts. However, in the same work, O'Loughlin notes that studies using a gridded approach often failed to implement rigorous spatial statistical analysis to account for the two spatial effects: spatial autocorrelation and spatial heterogeneity. Spatial dependence (or autocorrelation) refers to the tendency for nearby units to exhibit similar values, while spatial heterogeneity (or spatial nonstationarity) captures the idea that relationships between variables may vary across space rather than remaining constant. Without accounting for these spatial effects, analyses yield heavily biased estimates and obscure meaningful patterns, often misinterpreting null results and overlooking novel findings. Although spatial autocorrelation has trickled into the field, to the author's knowledge, there currently exist no papers in international relations that account for spatial heterogeneity using formal spatial statistical methods. While some studies do attempt to address heterogeneity by running separate models for each country, this approach assumes that country borders align with meaningful variation in the data—an assumption that often does not hold.

The small cohort of papers which account for spatial autocorrelation coupled with the pervasive lack of papers which account for spatial heterogeneity likely points to a deeper problem in the field. Without additionally accounting for spatial heterogeneity, many spatial regressions will produce null results. Further, we assume there exists a strong publication bias towards papers with meaningful results and that spatial heterogeneity is likely omnipresent in any political analysis with broad geographic scope – as is standard in International Relations. As such, it seems likely that there have been many promising projects stymied by their failure to move past null results by further modeling nonstationarity.

This paper demonstrates one such example where an analysis of spatial heterogeneity allows for the identification of a strongly significant region in an otherwise meaningless spatial model. I will utilize data from the Kurdish Conflict from 1980-2015, and perform analysis both with simple OLS and spatial regression models at the grid level and then utilize multiple regimes analyses. Alongside endogenous regimes, my primary tool for addressing spatial heterogeneity, this paper explores a range of other methods designed to capture spatial variation. By going through the most updated toolbox of spatial heterogeneity analysis techniques, I aim to convince the readers of their importance, while also providing a framework for their incorporation into future research. My hope is that these tools will both improve our understanding of how conflict shifts in space and also open up the possibility for an increased number of spatial regression models in conflict studies, which as such remains quite low.

First, this paper will discuss how spatial effects have been incorporated into Conflict Studies Research. The paper will then outline what these spatial effects are and what techniques currently exist to address them. The paper will then provide a case study using data from the Kurdish Conflict and some of the aforementioned spatial effects methodologies to demonstrate the importance of accounting for these effects.

CHAPTER 3

SPATIAL EFFECTS

In spatial econometrics, spatial dependence and spatial heterogeneity are the two fundamental spatial effects that require specialized methodology [2]. Though these effects have received considerably less attention in International Relations, outside the field, they are the source of extensive study and have inspired an ever-evolving set of methodological approaches. In this section, I will discuss how the field Conflict Studies has thus far accounted for these spatial effects and explain their mathematical foundations. I will also detail prominent methodological approaches for addressing these spatial effects and how these approaches may be applied to sub-national conflict data.

Accounting for spatial autocorrelation and spatial heterogeneity have become a subject of discussion in Conflict Studies work as sub-national data structures, especially grid based data structures have entered the literature. As has been pointed out by scholars like O'Loughlin and Raleigh [19, 24], grid based data structures, when deployed with statistically appropriate methods, are incredibly powerful tools in the study of intra-state conflicts. Primarily, grids can provide a more granular unit of analysis than countries or ethnic groups. Not only does this allow data to capture significant regional diversity across large territories, but it also allows researchers to choose the scale most approprite to their phenomenon of interest [23]. For conflict studies this is especially critical. In the study of conflict diffusion, for instance, small grid cells allow for a precise analysis of conflict's spread over time [20, 32]. Grid structures are also exogenous to the dynamics they model. In modeling social phenomena, it is inevitable that the variables as well as the boundaries which aggregate them have been influenced by the same social forces [29, 18]. For a uniformly applied grid, aggregation is comfortably exogenous to the system. Finally, grid data structures are incredibly useful when analysis crosses national boundaries. Political sub-national units in one country are not equivalent to those in another, so an exogenous grid is necessary for international comparisons.

In 2009, when O'Loughlin wrote his review of spatial analysis and conflict [19], there were only a handful of studies using a gridded approach and these studies failed to implement rigorous spatial statistical analysis to account for autocorrelation and spatial heterogeneity. Since then, gridded analyses have increasingly trickled into the literature. Some of these analyses are predictive or examine diffusion in space and time, often using clustering and simulation techniques [19, 30]. However, there exists a handful of papers that fit a regression model in which each grid cell is a unit of observation. Some notable examples exist in the conflict literature on natural resources and climate change [17, 13, 8].

Assessing this small cohort of papers, I find that researchers have increasingly accounted for O'Loughlin's primary critique, spatial autocorrelation. Although earlier papers merely account for rudimentary spatial lag by controlling for neighboring locations [14, 15, 18], there is a small, but growing number of papers utilizing spatial econometric approaches that account for model specifications with diverse combinations of autocorrelation, as well as spatial autoregression [17, 13, 8]. Models that account for autoregression are essential as they not only capture a spatial spillover effect, but additionally capture the feedback loop created between an observation and its neighbors. Spatial econometric models are also essential due to the diverse forms of autocorrelation prevalent in conflict models. While conflict in one location may lead to conflict in neighboring locations (spatial lag), so too might the covariates be spatially structured (take drought, for instance). Furthermore, social science models notoriously have modest explanatory power, meaning an allowance for error autocorrelation is potentially critical. Together, these models ward against severe bias. Given the benefits allotted by spatial econometric modeling, it is a missed opportunity that they are not more prevalent in conflict literature.

Even more unfortunate, to the author's knowledge, there have been no articles that address O'Loughlin's second point of concern – spatial heterogeneity. Spatial heterogeneity is also crucial to account for in the analysis of conflict data. Not only does it allow for a deeper understanding of a model's spatial dynamics – the ultimate goal in any spatial analysis – but models which do not account for nonstationarity often have misleading or irrational results [2]. When any paper employing a spatial regression fails to examine spatial heterogeneity, it forfeits valuable information that could deepen its analysis.

However, failure to account for spatial heterogeneity likely points to a deeper problem in the field. It is safe to assume that there exists a strong publication bias towards papers with meaningful results. Concurrently, it may also be assumed that spatial heterogeneity is omnipresent in any political analysis with broad geographic scope – as is standard in International Relations. As such it seems likely that there have been many promising projects stymied by their failure to move past null results by further modeling nonstationarity.

While some studies examine effects within individual countries as a robustness check, this approach often fails to fully address spatial heterogeneity. Although some variation may align with national borders, much of it does not. Conflict dynamics frequently cut across or emerge within state boundaries, making country-by-country models an inflexible and often inadequate approximation. While the case study at the end of this paper will demonstrate these points, it is first crucial to understand spatial dependence and spatial heterogeneity, along with the diverse toolbox of methods available to address them. The following sections provide a detailed description of the two spatial effects and associated methodologies, especially focusing on spatial heterogeneity, the central focus of this paper.

3.1 Spatial Dependence

Spatial dependence, one of many forms of cross-sectional dependence, occurs when there exists correlation between pairs of observations derived from their ordering in geographic space. When a model is fit to these observations, there may exist dependence in the outcome variable, dependence in the covariates, dependence in the errors, or dependence in any combination of the three. This dependence may even lead to feedback effects, wherein changes in one spatial unit affects other units, which in turn feeds back to the original unit. If left unaccounted for, dependence leads to a host of issues associated with model misspecification—most prominently autocorrelated errors, heteroskedasticity, and biased estimates. As such, a special set of methods is required to deal with spatial dependence and feedback effects [2].

Spatial Regressions do precisely this. Treating each observation as a spatial unit, either point or polygonal, spatial regressions function similarly to traditional regressions, while capturing different types of dependence using spatial weights. Spatial weights record the neighbor structure between observations as an $n \times n$ matrix W (where n is the number of observations) with elements w_{ij} for each pair of observations i and j. The weight is one when i and j are neighbors and zero otherwise. Neighbor structure may be based on contiguity, distance, or a variety of less popular criteria, providing flexibility for the modeling of different phenomena. In practice, weights are row-standardized with each binary weight divided by the row sum.

$$w_{ij}^{(s)} = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}}$$

Row standardization leads to the construction of a spatially lagged variable or spatial lag, as an average of neighboring values [2]. For an observation z_i of a variable at location i, the spatial lag is constructed as:

$$i = \sum_{j=1}^{n} w_{i,j}\, z_j,$$

Here, the weights w_{ij} consist of the elements of the i-th row of the weights matrix W, matched to the corresponding vector of observations on z. The spatial lag then becomes the weighted sum of the values observed at neighboring locations.

In spatial regression models, many specifications can be obtained by applying the spatial lag operator to different parts of the model, thus accounting for a diversity of spatial dependencies. The most common of these is the spatial lag model [1], a mixed regressive, autoregressive model. Expressed in matrix notation:

$$\mathbf{y'} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

y'[2] is an $n \times 1$ vector of observations on the dependent variable. W is a $n \times n$ spatial lag operator with Wy as the spatial lag term with an an autoregressive parameter ρ, X is an $n \times k$ matrix of observations of explanatory variables with a K*1 coefficient vector B, and e is a $n \times 1$ error vector. This model can be interpreted as capturing the simultaneous effects of the effects of the neighbors on y. This simultaneity is more obvious when this equation is further reduced.

Other common models are the SLX model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\gamma} + \mathbf{e},$$

where instead of a WY term, WXγ is used with γ as a vector of coefficients capturing the spatially lagged explanatory variable; and the spatial error model:

$$\mathbf{e} = \lambda \mathbf{W}\mathbf{e} + \mathbf{u}.$$

which incorporates a spatially constructed error term to a standard regression model. Here λ is an autoregressive parameter [3] with u as the remaining error effect. This model

1. The spatial lag model without a simultaneous autoregressive component is known as a SAR error model.

2. The spatial lag model includes the dependent variable on both sides to reflect spatial interdependence: y is influenced not just by covariates, but also by the outcomes of neighboring units (Wy)

3. Spatial Error Models can take on a variety of forms. The Autoregressive Form and the Spatial Moving Average form being the most popular.

specifically allows for dependence in the error structure, where unobserved shocks drive correlation between observations. While each of these models accounts for spatial dependence in the Ys, Xs, or errors, there are many models which account for combinations of spatial dependencies. A hugely popular example is the Spatial Durbin which incorporates both a spatial lag on the Y and the Xs. There are also a variety of more complex models incorporating higher order neighbors, moving average processes, and any number of other elements beyond the scope of this paper.

In Conflict Studies, the potential benefits of incorporating spatial regression modeling are huge. A central facet of conflict is that violence in one location is apt to ignite conflict in neighboring locations. Often called conflict diffusion or conflict contagion in the literature [32, 7], in regression modeling, this process is best accounted for through a simultaneous spatial lag. While earlier papers which attempted to account for conflict spillover merely account for rudimentary spatial lag by controlling for neighboring locations [14, 15, 18], this rudimentary spatial lag does not account for the feedback effect captured in a spatial lag model. This is problematic, as it is easy to imagine how conflict in a given location might ignite nearby violence, which in turn could perpetuate or intensify the original conflict.

Spatial regression modeling also has the flexibility to incorporate spatial dependencies beyond conflict spillover. In the conflict and natural resources literature, there have been a few notable papers which examine covariate dependence [17, 13, 8], which is unsurprising given the huge dependence in the spatial ordering of natural phenomena. Further, it would not be surprising to see an increasing number of papers which account for error autocorrelation. Given what little variation is often accounted for in social science models, it may be important to measure autocorrelation of unknown origin.

Despite the great benefits allotted by spatial models, there are relatively few in the conflict literature. Though lack of methodological exposure may account for some of this under representation, it is more likely that the distinctive nature of conflict's inferential infrastructure has plagued any attempt at spatial modeling with an unrelenting list of methodological difficulties. While the discussion at the end of this paper will more thoroughly expand on the long list of empirical difficulties inherent to these analyses, I intend to spend the majority of this paper focused on a primary culprit: spatial heterogeneity.

3.2 Spatial Heterogeneity

Spatial Heterogeneity refers to the geographic variation in relationships or processes. It creates problems in conflict analyses when it leads to nonstationarity, where the relationship between the outcomes and predictors is non constant over space. In standard spatial regression, we assume a homogeneous model, where all observations share the same model and regression coefficients:

$$y_i = \alpha + \mathbf{x}_i' \beta + \epsilon_i,$$

Here, looking at a standard OLS model, α and β are fixed across all observations. However, using specifications which model spatial heterogeneity, α, β, or any number of and combination of other model parameters may vary across space.

Methods which model this heterogeneity are traditionally split into two camps: continuous heterogeneity and discrete heterogeneity [3]. In the continuous approach, observations are allowed to vary smoothly over space. In a linear model, this means every observation has its own intercept and or slope, estimated using a different, spatially weighted subset of the data centered around that observation.

$$y_i = \alpha_i + \mathbf{x}_i' \beta_i + \epsilon_i,$$

This method is distinct from discrete heterogeneity, also known as a spatial regimes approach [2]. A spatial regime consists of a subset of the data, with each observation belonging to a single regime. For spatial regime formation, observations in a given regime are connected through the spatial network structure, such that observations in the same regime do not belong to geographically disconnected subsets. The regime's specification for a linear model is as follows.

$$y_{ij} = \alpha_j + \mathbf{x}'_{ij}\beta_j + \epsilon_{ij},$$

Here we see that each observation is indexed individually by i, but also by the regime to which it belongs j. In this model, there are J total spatial regimes (j = 1,. . . ., J), each with their own intercept and slope.

There are two types of spatial regimes: exogenous and endogenous. Exogenous spatial regimes are determined apriori and are based on some criteria relevant to the subject of study. Typically, these are politically delineated administrative units, such as municipalities or counties. However, exogenous regimes range significantly. For example, much of the current literature which utilizes spatial regimes looks at predefined housing markets [4]. Endogenous regimes on the other hand are regimes derived from the data itself, hence endogenous to the model. In an endogenous regimes approach, model estimation is carried out jointly with regime determination. As we will see, there also exists methods in between the endogenous and exogenous approaches.

In the following sections, I will first address the great importance of incorporating spatial heterogeneity analysis into any conflict-related spatial regression analysis. The second section will discuss in greater detail the plethora of methods which exist to address continuous and discrete spatial heterogeneity, while expanding on the existing methodological toolbox to propose some variations that, while less commonly used, provide necessary modeling flexibility. Through the exploration of this diverse set of methodologies, I hope to provide a useful road map for scholars conducting conflict studies who wish to advance their work by examining spatial heterogeneity.

3.2.1 Spatial Heterogeneity in Conflict Studies

To the author's knowledge, there are currently no papers in Conflict Studies that empirically examine spatial heterogeneity. Some papers use exogenous regimes as a robustness check for a global finding, and there exist some papers which examine continuous spatial heterogeneity for urban crime [1], but this accounts for the extent of the methods' application. Unlike the under-utilization of spatial regression, here there is a more clear culprit: the methods to account for continuous spatial heterogeneity and endogenous regimes are relatively new and as such have not made their way to the field.

Yet, there are compelling reasons the field should take up this methodology. International Relations tends to work with large geographic regions, often those spanning one or multiple countries. The scope of this investigation makes models ripe for spatial heterogeneity. Diverse geopolitical realities inevitably mean that what trend exists in one location may not exist in another. On one hand, even if a model can successfully be fit to a wide geographic area and provide meaningful results, there is never a downside of further investigating how these results vary across space. The researcher will still gain significant insight into the dynamics of their data – which is the ultimate goal of any project. On the other hand, models which do not account for nonstationarity often have misleading or irrational results [2]. Modeling globally may not produce results, where a more local analysis would. Further, heterogeneity often leads to coefficients that violate Tobler's first Law of Geography, which states that near things are more alike [27]. When lagged coefficients are greater in magnitude or have signs that do not match their non-lagged counterparts, this implies farther units have a greater impact on a given unit than nearer units. Rarely is this a fair assumption. Hence, it is worth exploring what methods exist to address spatial heterogeneity and which may be best suited for conflict studies.

3.2.2 Methods for Spatial Heterogeneity Analysis

Continuous Spatial Heterogeneity is addressed through Geographically Weighted Regression (GWR) [10] and Multiscale Geographically Weighted Regression (MGMR) [11]. Because GWR is programmed in ArcGIS it is currently the most popular approach. GWR is a local regression technique that allows regression coefficients to continuously vary across space. In GWR, the global regression model given by:

$$y_i = a_0 + \sum_k a_k x_{ik} + \epsilon_i.$$

Where y_i is the outcome at location i, a_o is the global intercept, a_k is the global coefficient on covariate k, and x_{ik} is the value of coefficient k at location i. This equation is then transformed such that the global parameters become locally estimated parameters:

$$y_i = a_0(u_i, v_i) + \sum_k a_k(u_i, v_i)\, x_{ik} + \epsilon_i,$$

Where (u_i, v_i) represents the coordinates of the ith point in space and $a_k(u_i, v_i)x_{ik}$ represents the coefficient of x_k at location i. These coefficients are estimated using a weights matrix at each i location, which uses a weights function, often a Spatial Kernel [4], to determine the weight given to each neighboring observation in estimating the ith coefficient. These kernels are typically distance based, assigning smaller weights to more distant observations. However, kernels are typically coupled with a bandwidth, fixed (distance based) or adaptive (KNN), which determines the cutoff for which observations are used to estimate coefficient i. These bandwidths are chosen using cross-validation (CV)[5] or the Akaike In-

4. A Spatial Kernel is a function used to calculate the density or influence given to spatial features.

5. Cross-validation is a machine learning technique which divides the data into multiple folds, using one of these folds at a time as a validation set, while training the model on the remaining folds. An average of these validation steps is used to create a more robust estimate of the model's performance.

formation Criterion (AIC)[6] and are applied to all the observations for each variable in the data.

This is the major point of contrast between MGWR and GWR. Unlike GWR, MGWR re-estimates a new bandwidth for every observation in the data. This is accomplished using an iterative back-fitting algorithm where each individual bandwidth is optimized to minimize the CV or AIC. Though MGWR is ideal for capturing more local effects which rapidly vary across space, it is also extremely computationally intensive. For this reason, unless there is a compelling theoretical reason to implement MGWR, GWR is often sufficient.

However, there are some downsides to GWR and MGWR that are especially relevant in conflict research. Depending on the spatial distribution of the data, some observations may be estimated using vastly different numbers of neighbors or with bandwidths of significantly varying sizes, resulting in highly uneven levels of precision across locations. While GWR and MGWR can calculate the p value for each estimate and map points with significant coefficient estimates, traditional implementations of the two continuous heterogeneity methods do not account for precision. Though this is starting to change, the fact that most available tools still fail to account for precision limits the scope of interpretation. Further, GWR and MGWR are only applied using simple OLS regressions. These approaches do not have the flexibility to incorporate spatial regression models nor other complex relationships. As such, continuous spatial heterogeneity tools have a more limited scope in conflict research. Instead of making sweeping conclusions based on these models, which do not account for precision and spatial effects, analysis with GWR and MGWR should be primarily exploratory. These methods produce visualizations ideal for investigating how coefficients vary across space. Given their relative ease of implementation, they provide a fast way for a researcher to more deeply understand the dynamics of their data.

Yet, GWR and MGWR can also be used to address discrete spatial heterogeneity. This

6. The AIC estimates the quality of each model by estimating the amount of information lost, while balancing the tradeoff between model simplicity and overfitting.

has been accomplished using PCA[7], SKATER[8] [16], and more recently K-means[9] [12], to cluster coefficients from M/GWR and create distinct spatial regimes, an example of methods which fall somewhere between endogenous and exogenous regime formation [10]. These regimes may be taken and interpreted as is, or remodeled using another non-OLS analysis. Though this approach improves interpretability and in the cases where the model is refit, improves coefficient precision, a number of drawbacks remain. Firstly, M/GWR assumes a smooth distribution of coefficients, which is antithetical to a discrete model. This means there exists a conceptual mismatch unless the regime boundaries are fuzzy. Furthermore, the inflexibility of the approach carries over. Though spatial effects could be fit ex-post to the regimes, the original clustering must utilize basic OLS derived coefficients.

These problems are rectified by another formulation of endogenous regimes, spatially constrained endogenous regimes [4]. This regime estimation method is based on an extension of the SKATER clustering algorithm. This approach utilizes a graph partitioning logic which enforces spatial contiguity, where only nodes (observations) that are spatial neighbors have an edge between them in the graph. This graph itself is structured using the $n \times n$ spatial weights matrix, where each observation is a node and each edge corresponds to a non-zero element in the spatial weights matrix, the sum of which make up the neighbor structure among the observations. The edge weight is based on the attribute similarly between a pair of observations, computed by Euclidean distance. The resulting weighted graph is then reduced to a Minimum Spanning Tree (MST), such that there are n-1 edges connecting n nodes, and the sum of the edge weights is minimized. Ultimately, this MST will be divided

7. Principle Component Analysis (PCA) is a dimension reduction technique which reduces the the number of variables in a set, while maintaining maximum variation between the sets.

8. SKATER is a constrained spatial regionalization algorithm based on spanning tree pruning. The number of edges is pre-specified to be cut in a continuous tree to group spatial units into contiguous regions.

9. K-means is an iterative, centroid-based clustering algorithm that partitions the dataset into similar groups based on the distance between their centroids.

10. By using M/GWR estimates to create clusters which are then refit to spatial models, the user essentially creates exogenous regimes defined by the data

into regimes by fitting a regression to each regime based on the optimization of some criteria. For OLS and Spatial Lag models this is the Sum of Squared Residuals (SSR), computed by calculating the difference between the observed and predicted values.

$$\text{SSR} = \sum_i (y_i - \hat{y}_i)^2 = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}),$$

For each tree T, the removal of an edge E, results in two subtrees, a and b. The change in SSR that results from any given cut is computed as:

$$f(E) = \text{SSR}_T - (\text{SSR}_a + \text{SSR}_b)\,,$$

Where SSR_a and SSR_b are the sum of squared residuals (SSR) for each subtree. The edge E that is eventually pruned is the one for which the change in the $f(E)$ produces the largest reduction in the SSR. This cut produces one new regime. Hence, the process of fitting the regression model and then pruning the edge with the largest $f(E)$ continues until all J desired regimes are created.

This approach has some deficits. It is computationally demanding and because it uses a hierarchical method of pruning, once an observation has been assigned to one regime, it cannot be reassigned to another. Furthermore there are certain limits on the flexibility of applying different regression models. Current packages only account for OLS and spatial lag regression.

This being said, theoretically, spatially constrained endogenous regimes could account for any number of regression models. SLX and any model which can utilize a SSR based trimming criteria are compatible. Spatial Error models, which are incompatible with SSRs due to the intentional fixing of $\mathbb{E}[e \mid X] = 0$, could nevertheless use the model's AICs, though this would be highly computationally demanding due to the necessity of a Maximum Likelihood Estimation (MLE) based approach. A simple way around updating the existing packages to include new regression types is similar to some approaches we've already discussed: Run the SKATER algorithm, create regimes based on a simple OLS classification, and then re-model each regime using another more complex regression model. Ultimately, this is very similar to how exogenous regime models are run. While an intercept or slope may be held fixed across the models, this method essentially re-runs the model or each ad hoc regime. However, in this case, the ad hoc regimes were originally data derived.

Despite some drawbacks, spatially constrained endogenous regimes provide an excellent way to account for spatial heterogeneity in conflict data. Unlike M/GWR, they are more statistically robust. Endogenous regime approaches estimate fewer parameters while also producing much more robust inference. You can, for example, see both the standard errors for each regime model, the p values for each coefficient estimate, and run a chow test[11] across all the models. Furthermore, this approach has the flexibility to work with a diversity of spatial models covering a range of spatial dependencies. Even if the spatial models need to be fit following an OLS SKATER classification procedure, this may still be preferable to fitting spatial models to clustered GWR regimes, which derive clusters from inherently smooth continuous coefficients. Finally, unlike exogenously defined regimes, endogenous regimes do not have the burden of prior assumptions. Even if the produced regime turns out nearly identical to the exogenous regime, there is no way to confirm this relationship until the clustering algorithm is run. As we will see later in this paper, there are examples where a phenomenon is not bound by national borders. In ethnic conflict spanning multiple countries this is especially true. Hence, it is worth having a method with the flexibility to model at any level of aggregation.

11. The Chow Test is a statistical test used to determine whether the true coefficients of two or more tests are equal.

CHAPTER 4

KURDISH CONFLICT CASE STUDY

To demonstrate the insights that can be gained from incorporating heterogeneity analysis, I present a case study of violent events involving Kurdish populations from 1980 to 2015. Specifically, I model the relationship between conflict incidence and the spatial distribution of oil deposits. This case study was deliberately chosen for two reasons: first, to illustrate both the necessity and potential analytical power of incorporating spatial heterogeneity; and second, to highlight the methodological challenges that may arise when doing so.

This dataset is ripe for rampant spatial heterogeneity. The Kurdish conflict spans multiple countries, and both between and within these countries, the conflict is shaped by differing political, social, and economic dynamics. Accordingly, it was expected that different functional relationships would need to be fit across space. In line with the core aim of this paper, accounting for spatial heterogeneity in this context reveals results that would have remained hidden under a homogeneous model. This case study was also constructed to illustrate the challenges of working with gridded data that span both space and time. When data vary across both dimensions, spatial heterogeneity is often accompanied by temporal heterogeneity. The relationship between oil and conflict may not only differ across regions but may also evolve over time. This problem becomes especially difficult to address when the temporal variation is non-linear, though this study proposes and explores some viable solutions.

Further, oil was intentionally selected as the focal variable not only because it significantly overlaps with Kurdish territories and shapes the dynamics of the conflict, but also because it holds a prominent place in international relations literature—particularly in debates surrounding the so-called "resource curse." As such, readers are likely to be somewhat familiar with the non-methodological theory motivating this case study.

While the topic of oil and conflict may be familiar, this case study also exemplifies the broader potential of spatial heterogeneity analysis in conflict research. Today, data availability is sufficient to apply gridded-data approaches to many regions worldwide for most time periods after 1950. Although oil and other natural resources are classic topics for sub-national spatial regression modeling, one can easily envision a wide range of spatially oriented questions that would benefit from the methods presented in this paper.

4.1 Data

The data from this project are taken from the UCDP GED[1] [26] which provides geocoded point data on conflict events going back to 1980. All conflict events in Turkey, Syria, Iraq, and Iran whose descriptors included the word "Kurd" were extracted for use in this analysis. All other data was sourced from PRIO-GRID [2][28] which provided a .5 by .5 decimal degree grid and the other pre aggregated variables used in the analyses. The outcome variable, Kurdish Conflict Rate, is calculated by counting the number of conflict events in each grid square and dividing by the grid population. The primary explanatory variable is an indicator for oil presence, sensitive to year of discovery[3]. The model covariates are as follows: Urban land (ISAM-HYDE[4]), which provides the percent urban land in each grid square; Population, total (GPW data[5]); Distance to nearest Capital; and Distance to nearest Border. These covariates were selected because they are among the most commonly used controls in subnational conflict models examining the relationship between conflict and a natural resource.

1. Uppsala University Data Portal Georeferenced Event Data
2. Peace Research Institute of Oslo Grid Database
3. Ex. If an oil deposit is discovered in 2000, it will not be considered present in the 1999 data.
4. Global Historical Land-Cover Change and Land-Use Conversions Data
5. Gridded Population of the World Data from NASA

4.2 Background

Kurdistan is an ethnographic region spanning modern day Turkey, Iraq, Syria, and Iran. Following the collapse of the Ottoman Empire, in 1916, Kurdistan was divided into these numerous states, where the Kurdish people became ethnic minorities. Since, the Kurds have since faced varying efforts of forced assimilation and oppression. This oppression, while multifaceted, has been driven in part by regional oil deposits, the main predictor variable in this case study. Kurdistan overlaps with many oil-rich regions and national governments have in many cases sought to control these resources, leading to forced migrations and hostile relations with Kurdish communities.

During periods of relative regional stability, these tensions have largely remained conflicts between Kurdish populations and the specific states where they reside. However, in times of deeper unrest, these localized struggles have frequently escalated into broader geopolitical issues involving neighboring states as well as the international community. The presence of valuable oil resources has contributed to this dynamic, as external actors—including the United States and other foreign powers—have intervened to secure and protect oil-rich regions. This external involvement only underscores the diverse impact oil deposits have in shaping Kurdish conflict.

Though these circumstances have evolved over the past century, tensions between majority-ethnic governments and the Kurdish minority have given rise to numerous militant groups across Kurdistan, as ethnic repression has fueled both nationalism and armed resistance. Structurally complex, the conflict has ranged from localized protests to full-scale civil war, with shifting alliances among Kurdish factions, non-Kurdish insurgent groups, and national governments. This ongoing and multifaceted struggle for greater autonomy is commonly referred to as the "Kurdish Conflict." In order to obtain an overview of the Kurdish Conflict, it is important to examine Turkey, Syria, Iran, and Iraq individually. Though the movement crosses borders, unique political realities have nevertheless substantially shaped the conflict.

4.2.1 Turkey

With the creation of the Kurdistan Workers Party (PKK) in 1980, Kurdish Nationalism transformed from an unarmed struggle to one with violence central to political operation. Though swiftly dismantling opposition, and recruiting thousands from the countryside, the PKK's ascension stopped as quickly as it started. Following the coup in 1980, all Kurdish organizations were decimated. With many imprisoned and many others choosing to flee the country, it wasn't until 1984 that the fight for an autonomous Kurdistan resumed [6].

Led by Abdullah Ocalan, the PKK began to engage in a relentless campaign of guerilla warfare. Tens of thousands of militants joined their struggle, quickly attracting the wrath of the government. In 1989, Turkey began evacuating and burning Kurdish villages suspected of harboring PKK. This repression mobilized large swaths of Kurdish society, turning the PKK into a vast 'party complex' throughout the early 90s. The creation of legal groups, newspapers, and NGOs all helped fuel the ongoing conflict [6].

By 1993 the Turkish government began to reconsider its repressive tactics. Opting for increased instruments of recognition, the government and the PKK were able to broker a ceasefire. Following the capture of Ocalan in 1999, the era of rampant violence had come to a close. Accordingly, the subsequent years brought an ideological shift in the party. No longer seeking an autonomous Kurdistan, they instead advocated for democratic confederalism. Unrealized, this aim drove both a resurgence of violence in 2004, but also the founding of political parties which sought to advance Kurdish and leftist ideals. From 2004 to 2015 violence started to wane as political influence grew. Starting in 2009, negotiations for the disarmament of the PKK began. Unfortunately, political differences proved impossible to resolve and the agreement broke down in 2015. With this breakdown of negotiation, clashes between the Turkish state and PKK resumed, ushering in a new era of violence. By 2016, the PKK had been defeated, though political branches of the Kurdish movement continued to assert their influence [6].

4.2.2 Syria

Throughout the 1950s and 60s, the Kurdish population in Syria endured institutionalized discrimination, including waves of arabization and cultural suppression. This began to shift with the rise of Hafez al-Assad and the Ba'ath Party in the late 1960s. However, Kurdish political fragmentation limited the consolidation of power. Ideological divisions between progressive and conservative factions prevented effective unification, even as regional actors gained influence [6].

To counter growing Sunni Arab opposition, Assad sought limited ties with Kurdish Alawites in the northwest, offering religious and military positions to integrate them into his regime. This uneven co-optation both challenged Syria's neighbors (Turkey and Iraq) and preempted strong Kurdish mobilization within Syria. While the "Kurdish question" lay dormant domestically, Syria became a haven for PKK militants in the 1980s and 1990s, until pressure from a Turkish-Israeli alliance in 1996 led Assad to withdraw his support [6].

Following this period, Kurdish influence in Syria steadily grew. The Democratic Union Party (PYD) was formed in 2003, and after the "Damascus Spring" and the fall of Saddam Hussein, Kurdish parties briefly united. A 2004 riot in Qamishli following a football match revealed growing tensions, resulting in deaths and widespread demonstrations. Though the movement fragmented again, it laid the foundation for future Kurdish mobilization. By 2011, there were more than 20 Kurdish parties, though many were internally divided and suppressed by the regime [6].

The 2011 Arab Spring reignited Kurdish momentum. As the Assad regime withdrew from northern areas in 2012, PKK-linked cadres established the People's Protection Units (YPG) alongside the PYD. Though the Erbil Agreement sought to unify the PYD and the Kurdish National Council (KNC), the PYD-YPG quickly emerged dominant, expanding territorial control through both coercion and local alliances. Between 2012 and 2014, they captured large swaths of northern Syria, including key oil fields—an advance bolstered by

US military support during the fight against ISIS. Despite Turkish objections, this alliance allowed Syrian Kurds to secure roughly 25 percent of the country, a position they largely maintain today [31].

4.2.3 Iran

Following the Iranian Revolution in 1979, the Kurdish question in Iran can be split into four different time periods. The first period starts in the aftermath of the revolution and runs into the reformist era of the 1990s. Characterized by widespread violence, during the consolidation of the new government, Kurds and other ethnic minorities were largely seen as a security threat. Armed groups were deployed in Kurdish regions to intimidate local populations and carry out widespread, random acts of violence. This reality did not change following the Iran-Iraq war in 1988. By 1990, 25 percent of Iranian troops were stationed in Kurdistan and maintained tight control over the region. The primary mode of Kurdish resistance during this era was a sustained campaign of guerilla warfare [6].

Starting in the 1997s, Muhammed Khatami came to power, advancing a reformist agenda which included support for minorities. Khatami aimed to reshape government and civil society, partially by reintegrating a Kurdish polity. Kurdish people were now allowed some degree of political expression and could advocate for change. Though more Kurdish leaders were integrated into government and many civil rights were returned, the reform government failed to implement the great majority of its promises [6].

The third time period saw the return of hardliners to government. Mahoud Ahmedinejad won the presidency in 2005 championing populist rhetoric and a security oriented approach to minority populations. Equal citizenship rights were replaced by a militarized political system. Ahmedinejad's government intensified politically motivated prosecution, censorship, and many forms of political repression. Starting from this time period, guerilla warfare against government forces resumed [6].

The forth time period starts with the 2013 election of Hassan Rouhani. Though Rouhani ran on a platform of advocating for religious minorities, little changed. Once in office he appointed the former minister of security and intelligence, marginally increased acceptance of Kurdish language learning, and appointed only a handful of Irani Kurds to middle level administration. As such the conflict between the government and Kurds persists [6].

4.2.4 Iraq

In Iraq, the discovery of the Kirkuk oil fields in 1927 marked the beginning of systematic Kurdish suppression. Successive governments in Baghdad marginalized Kurdish populations to prevent their participation in oil production and revenue sharing. This dynamic fluctuated with the strength of the state — concessions were offered when Baghdad was weak and retracted when it regained strength [6]. By the late 1960s, Kurdish control had expanded in northern Iraq, prompting the government to launch a counter-campaign. A 1974 peace attempt by then–Vice President Saddam Hussein failed over the disputed status of Kirkuk, leading to renewed state control and intensified conflict [6]

The latter 1970s brought the second Iraqi-Kurdish war and a campaign of Arabization that forcibly displaced Kurds from oil-rich regions. With the outbreak of the Iran-Iraq War in 1980, Iran supported Kurdish factions, and in turn, Kurdish backing of Iran provoked harsher repression. By 1986, the Anfal Campaign had begun, leading to the mass killing of Iraqi Kurds and one of the most severe episodes of ethnic violence in the region's history [5].

Following the 1990 Gulf War, the UN imposed a no-fly zone over northern Iraq, effectively shielding Kurdish populations from air strikes and enabling the formation of a de facto autonomous region. The Kurdish Regional Government (KRG) emerged in 1992, though its authority was contested by rival Kurdish parties—the KDP and PUK—whose four-year civil war ultimately resulted in a dual-power structure in the north. When Saddam Hussein's regime fell in 2003, the KRG quickly became the most stable and experienced governing entity in Iraq. It secured recognition as a semi-autonomous federal region with control over its own oil resources, and by 2007, Turkey supported the export of KRG oil through the Kirkuk-Ceyhan pipeline [6].

Despite this growing autonomy, tensions with Baghdad escalated under Prime Minister Nouri al-Maliki. A potential civil war was averted only by the rise of ISIS in 2014. The KRG took on significant military and humanitarian responsibilities, expanding its control over key oil regions like Kirkuk. While this expansion was later reversed under international pressure, the KRG retained its semi-autonomous status—made possible by its persistent control over oil production and its strategic value in regional security [6].

4.2.5 Hypotheses

Given this background, several hypotheses were developed prior to the data analysis stage. Drawing on existing findings in the literature, I first anticipated a positive and significant relationship between conflict for both the percentage of urban land and distance from the capital and the border, reflecting common findings in subnational conflict models. Also found in prior literature, I expected most of the models to exhibit a spatial lag, indicative of conflict spillover. However, due to the diverse political and spatial dynamics across the four countries in the study, I did not expect pooled models—those aggregating all countries together—to yield significant results for oil presence. Instead, I hypothesized that country specific and endogenous regimes models would better capture the relationship between oil and conflict.

In Iraq, I expected a positive association between oil presence and conflict after 2003, driven by territorial disputes that emerged following the fall of Saddam Hussein. In Syria, during the civil war period, I anticipated a negative relationship, as U.S. military presence helped secure oil-rich areas and Kurdish forces increasingly consolidated control over these regions, particularly toward the end of the conflict. For Turkey, I did not expect oil presence to be statistically significant in the exogenous regime analysis. However, I predicted that its eastern regions would merge spatially with parts of Syria in the endogenous regime analysis, particularly for the first (1997–2003) and fourth (2011–2015) models, due to cross-border PKK activity and Turkey's military engagement in the Syrian Civil War. In the first period, I did not expect oil to be significant; but in the fourth, I hypothesized a significant relationship similar to that observed in Syria, primarily driven by US securement of oil.

Finally, for Iran, due to the scarcity of conflict event data, I did not expect to observe any meaningful or statistically significant relationships in this region.

4.3 Analysis

Clearly, the Kurdish populations in these four countries have experienced a significant degree of political change from 1980-2015. As such, any model which pools these years together is unlikely to derive meaningful results. Though it is quite common to employ spatial analysis coupled with longitudinal analysis when handling temporally expansive spatial models, I have instead chosen to divide the data into groupings of years with cutoffs between major political events. The data was divided into the years before and after the overthrow of Saddam Hussein (1997-2003 vs. 2004-2009) and before and during the Syrian Civil War (2006-2010 vs. 2011-2015). These divisions aid data interpretability and are preferred over the incorporation of a time series approach which is notoriously difficult to interpret in-tandem with spatial models[6].

For each of the four data groups, I conducted a preliminary OLS analysis. Because Iran exhibits substantial zero inflation due to limited conflict activity, I estimated two models: one including Iran and the primary reported model excluding it. Given there is significant

6. Time series approaches pose significant challenges for spatial modeling, particularly when applied over long periods. Temporal nonstationarity makes it difficult to specify models that remain stable and interpretable across time. When combined with spatial dependence and heterogeneity, these models become highly complex and often obscure the distinction between spatial and temporal effects[9]. As such, dividing the data into politically meaningful periods offers a more interpretable framework

theoretical reason to believe these models may exhibit spatial effects, I then ran spatial specification tests. It is likely that conflict in one location would have an effect on conflict in another location. This spatial lag could be motivated by grievance or the tendency for violence to diffuse away from strongholds of an insurgency. Further, spatially structured data for such a rudimentary model may certainly have spatially dependent unobservables best captured by spatial error dependence. Following a round of specification tests, these theoretical concerns seemed to be well founded. I fit three spatial lag models and one spatial error model (2011-2015), all using an Instrumental Variables (IV) approach, with a first order spatial lag and queen's contiguity one weights.

In my final stages of analysis I compared an exogenous regimes approach, wherein I re-examined the results from the other two results for each of the individual four countries in my model; and an endogenous regimes approach, where I re-examined my OLS models using data driven regime classifications. The number of regime classifications was determined by Mojena's rule. Though I additionally attempted to incorporate GWR and fit spatial models to my exogenous and endogenous regimes, I was unable to do so given limitations in my model. Hence, for the three most meaningful regimes derived from my spatially constrained endogenous regimes analysis, I re-ran them using an OLS specification and after ran a specification search to fit them to respective spatial models, using the same IV approach and weights classification.

4.4 Results

My first set of results comes from a standard OLS model for each separate time period (1997-2003; 2004-2009; 2006-2010; 2011-2015). Owing to the severely zero inflated Iranian data, I present these OLS results in Table 4.1 without this data. While the OLS results which include all four countries do not yield much of significance, there is some information to be gleaned once Iran is removed. Across these simple regressions, there are some consistent results. Distance to the Nearest Capital and Distance to the Nearest Border are almost always positive and significant, with the exception of 2011-2015 in which distance to the border has no significant effect on Kurdish Conflict Rate and Distance to the Capital has a negative effect. However, these OLS results lack significant explanatory power. No model exceeds an Adjusted R Squared above .1, hence minimal variance is captured by coefficients in the models.

Variable	**Model 1**	**Model 2**	**Model 3**	**Model 4**
CONSTANT	-0.00844***	-0.00507***	-0.00418***	0.02274**
	(0.00000)	(0.00059)	(0.00051)	(0.00657)
petroleum_y	0.00131	0.00113	0.00038	0.00851
	(0.31101)	(0.74083)	(0.68362)	(0.18617)
log_urban_ih_filled	-0.00008	-0.00012	-0.00012	0.01937**
	(0.95455)	(0.95116)	(0.90328)	(0.00481)
bdist3	0.00001***	0.00001**	0.00002**	-0.00006
	(0.00044)	(0.00334)	(0.00605)	(0.19852)
capdist	0.00002***	0.00001***	0.00001***	-0.00004**
	(0.00000)	(0.00000)	(0.00000)	(0.00937)
R-squared	0.0804	0.0423	0.0398	0.0276
Adjusted R-squared	0.0745	0.0365	0.0398	0.0217

Table 4.1: Comparison of four OLS regression models
Note: P Values in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

My second set of results comes from the addition of spatial effects in the models (Table 4.2). Models 1, 2 and 3 (1997-2003; 2004-2009; 2006-2010) are appropriate for a spatial lag model and all three exhibit significant spatial lag effects with coefficients around .8. Model 4 (2011-2015) instead exhibits spatial error dependency, with a significant lambda and a coefficient of .4. Though these spatial effects prove impactful, their inclusion also renders the other variables in the model insignificant. Though the R squared increases for every model, these R squares are still relatively low (all but one remains below .1), and the majority of the explanatory power comes directly from the spatial effects.

The third set of results is from my exogenous regimes analysis. In this analysis each

Variable	Spatial Lag 1	Spatial Lag 2	Spatial Lag 3	Spatial Error
CONSTANT	-0.00193	-0.00119	-0.00107	0.01874
	(0.40832)	(0.56056)	(0.51488)	(0.09105)
petroleum_y	-0.00035	0.00017	0.00019	0.00722
	(0.77821)	(0.87513)	(0.83058)	(0.48624)
log_urban_ih_filled	-0.00005	-0.00021	-0.00021	0.02093
	(0.69169)	(0.85901)	(0.83261)	(0.10116)
bdist3	0.00000	0.00001	0.00001	-0.00004
	(0.42400)	(0.60523)	(0.59422)	(0.40191)
capdist	0.00000	0.00000	0.00000	-0.00003
	(0.38092)	(0.55538)	(0.49500)	(0.09356)
W_log_conflict_rate	0.84313***	0.85668**	0.83713**	–
	(0.00014)	(0.00807)	(0.00653)	
lambda	–	–	–	0.40008**
				(0.00209)
Pseudo R-squared	0.1048	0.0938	0.0952	0.0272

Table 4.2: Comparison of spatial regression models: three spatial lag models and one spatial error model. Coefficients shown with p-values in parentheses.
Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

grid square was divided into one of four countries: Syria, Iran, Iraq, and Turkey. These individual country results for each of the four models yielded results similar to those from my simple OLS analysis. Distance to the nearest Border and Capital, and to a lesser extent the percentage of urban landscape fluctuated in significance, all with models which consistently captured little variation. Owing to the weakness of the models, I was unable to couple a spatial analysis with the exogenous regimes analysis.

The fourth set of results comes from the endogenous regimes analysis. For each of my four models, I created a set of endogenous regimes. These results are summarized below. While there exists a separate regression model for every separate data driven regime for each of the four models, for almost every model there is at most one regime containing results of any significance. Model 1 (1997-2003), as demonstrated by the Chow Test (Appendix Table A.1), has largely similar results to the other analyses. Distance to the nearest capital and border fluctuate in their significance across the regimes. However, for the other three models, one regime stands out from each as highly relevant for additional consideration. For

Model 2 (2004-2009), Regime 2 has a significantly higher Adjusted R Squared at .4 and has strongly significant positive coefficients, with the exception of distance to the nearest border. For Model 3 (2006-2010), Regime 3 has a higher Adjusted R Squared at .13 and has strongly significant positive coefficients, though for the two distance measures the effects are negligible. Finally, for Model 4 (2011-2015), Regime 2 has a higher Adjusted R Squared at .25 and has some significant coefficients. The effect of urban land cover is positive and strongly significant, while the effects of distance to the capital is somewhat significant and negative. All these results are summarized in table 4.3 and the regimes are viewable in Figure 4.1.

Variable	**Regime 2 (2004–2009)**	**Regime 3 (2006–2010)**	**Regime 2 (2011–2015)**
CONSTANT	-0.00167***	-0.00076***	0.06790*
	(0.00000)	(0.00010)	(0.01584)
petroleum_y	0.00499***	0.00556**	-0.06928
	(0.00000)	(0.00355)	(0.23354)
log_urban_ih_filled	0.00091***	0.00039***	0.10426***
	(0.00100)	(0.00065)	(0.00000)
bdist3	0.00000	0.00000*	-0.00206
	(0.06188)	(0.01691)	(0.13025)
capdist	0.00000***	0.00000***	-0.00013*
	(0.00088)	(0.00012)	(0.04876)
R-squared	0.4727	0.1402	0.2825

Table 4.3: Endogenous Regimes Results: Display the regime for each model which contains the most significant results. P-values in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Owing to the weakness of the models I was not able to run GWR, the exogenous regimes analysis or the endogenous regimes analysis using a spatial specification. While the regimes from the exogenous regimes analysis remained too weak to fit a meaningful spatial specification, some of the models from the endogenous regimes analysis proved sufficient. My final set of results come from refitting a spatial model to Cluster 2 from the 2004-2009 data, cluster 3 from the 2006-2010 data, and cluster 2 front the 2011-2015 data. For cluster 2 from

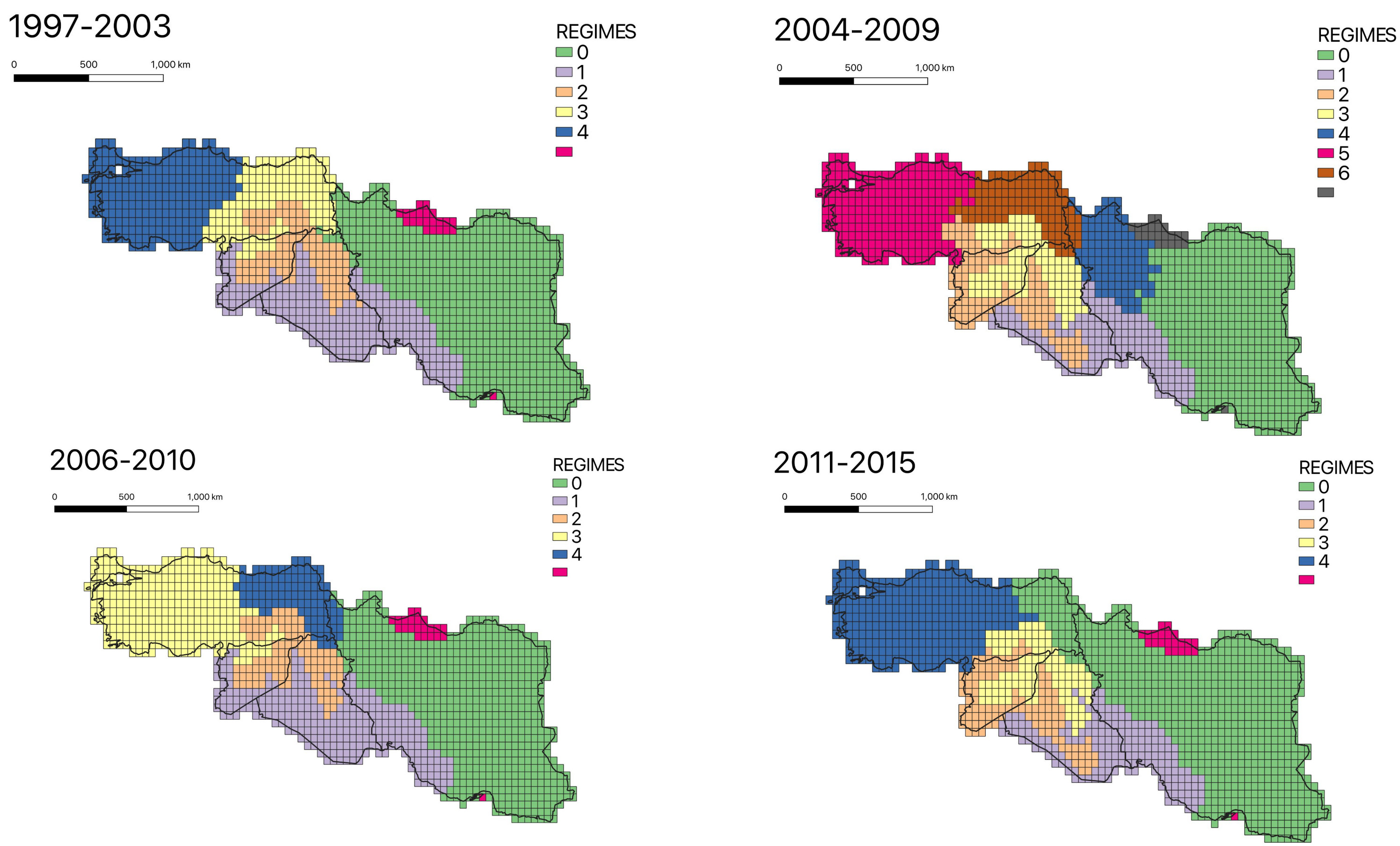

Figure 4.1: Figure 1: Kurdistan Spatial Constrained Endogenous Regimes Analysis

2004-2009 data, I find a positive spatial lag effect of .3 with a significance level of about .8. The other variables in this model retain their characteristics from their OLS counterpart, but the R squared raises to about from .47 to .57. Cluster 3 from the 2006-2010 data is less conclusive. While the specification tests hover between an error model, SDM model, and an SLX error model, only the SDM model produces results similar to those of the OLS counterpart—incorporating a lag for petroleum but yielding a very similar R squared. Cluster 2 from the 2011-2015 data does not seem to exhibit any spatial effects. As the result for cluster 2 from the 2004-2009 is most meaningful, a table with the results for this spatial analysis is included in the appendix (Appendix Table A.2).

CHAPTER 5

DISCUSSION

The results illustrate the two primary points of this paper. Firstly, that it is potentially critical to utilize spatial econometric modeling. Secondly, that prior to the introduction of endogenous regimes, an analysis (spatial or otherwise) may yield little of note. Starting off the analyses, an OLS for the entire pooled dataset produced nothing of significance. Even with Iran removed, the new OLS model achieves little. With extremely small R Squares, even for the Social Sciences, they can hardly even reaffirm what is already known in the field, that there exists an effect for distance to the nearest capital and border. Fitting spatial models only exacerbates the nullity of the covariates as the spatial effects render everything else insignificant.

This being said, the fact that this analysis yielded strong spatial effects does support the importance of using spatial models. The spatial error model especially points to this fact. It is certainly important to capture conflict rate spillover, but as we have seen there are a few potential ways to accomplish this. Yet, it is critical to utilize models which have the flexibility to capture autoregression and diverse forms of spatial autocorrelation – not just lag. Without accounting for spatial error, model 4 would have suffered from a severe case of correlated residuals, further biasing an already problematic model.

Following the OLS analysis, I divided my data into exogenous regimes (in this case country boundaries), which also did not reveal any significant relationships in the model. If I were to end my analysis here, following the convention in conflict studies, I would walk away maligning a poorly constructed model which yielded no results of interest. Luckily, I went on to implement spatially constrained endogenous regimes.

Endogenous regimes yielded a few results of interest. Firstly, for three out of the four model analyses, there existed a regime with great explanatory power. This highlights the benefit of investigating spatial heterogeneity. While the entire space cannot be fit to a single model, subsets of the space exhibit important spatial patterns. The second outcome of interest is the consistency of the regimes over time and that these regimes did not follow exogenous political boundaries. While many of the regimes were insignificant, the same shapes seemed to persist and these shapes did not follow existent political boundaries. As previously discussed, in International Relations and Conflict Studies, country boundaries are often the unit of aggregation or the bounding box for analysis. If there are latent spatial structures which do not conform to these boundaries, it begs the question of what the field overlooks by assuming that political phenomena occur within countries or at established subnational administrative borders. Data driven regimes seem to provide the flexibility to test the assumption that a phenomenon is sufficiently satisfied by exogenous regimes, and then if it is not, to uncover what spatial structures do exist in the data.

These findings offer qualified support for the paper's initial hypotheses. While the expected relationships between conflict and both distance to the capital and border were consistently significant, the role of oil and urban land proved more context-dependent, varying across space and time. Crucially, and somewhat unexpectedly, the inability of both pooled and exogenous-regime models to detect these patterns reinforces the central claim: that meaningful spatial variation is often masked when spatial heterogeneity is not explicitly modeled. By uncovering localized relationships through endogenous regimes, this analysis reveals conflict dynamics that would otherwise remain hidden. The following section turns to a deeper interpretation of these spatial configurations and what they suggest about the evolving geography of the Kurdish conflict.

5.1 Regime Interpretation

What do the endogenous regimes for each model tell us about the Kurdish Conflict? While this question is impossible to definitively answer, information about the historical progression of the Kurdish Conflict and its geography provide for ample speculation. Firstly, we can interpret regime 2 in Model 2. As this Model covers 2004-2009, this time period takes place after the fall of Saddam Hussein in Iraq and sees the rise of the Kurdistan Regional Government (KRG). While the regime is largely outside the KRG, it does cover contested territory holding significant oil deposits. As the KRG was only allowed to start selling oil in 2009, it makes sense that this time period was characterized by conflict surrounding oil ownership. Hence, it is likely that as the KRG sought to expand, the parts of Regime 2 in Iraq may have been contested specifically for their oil resources. The fact that this model requires a spatial lag for conflict is entirely consistent with theorizing done earlier in this paper on conflict diffusion.

Meanwhile, conflict dynamics in Turkey and Syria were also evolving in ways that help explain the spatial structure of Regimes 2 and 3. 2004 also saw the breakdown of the ceasefire between the Turkish Government and the PKK. While there were increased protests in urban cities like Diyarbakir, many PKK fighters moved towards the Syrian/Turkish border to enable movement between the two countries. This would explain the significant part of the regime that lies along the Turkish/Syrian border. This same development can likely be attributed to Regime 3 in Model 3 (2006-2010), which spanned the same conflict period, which did not end until 2012. While this partially explains the spatial distribution of Regimes 2 and 3 in these two countries, there is much unexplained. Specifically, it is unclear why oil might affect conflict during this time in Turkey and Syria.

Interpreting Regime 2 in Model 4 (2011-2015) is a bit more straightforward. As this time period covers the first part of the Syrian War along with the fights against terrorist groups like Al- Qaeda and ISIS, the strong significance of the percent of urban land cover likely points to defensive and offensive actions taken by Kurdish fighters around this time. Kurdish military groups across Syria and Iraq both secured their territory, especially defending strategic urban areas, while also assisting other groups to defend cities outside of the Kurdish purview.

5.2 Empirical Considerations and Drawbacks

Though this analysis produced some exciting results, the approach yields many methodological concerns. By discussing the challenges of conducting spatial analysis, both regression and regimes based on conflict point data, I hope to illustrate how future analysis may improve on my own research, but also allow readers to recognize which questions are most suited to the analytic approach described in this paper. Spatial regression analysis and spatial regimes analysis hold significant latent value in the field of conflict studies, yet different constellations of the two, and often neither, are appropriate for certain questions.

From the offset, available conflict point data creates difficulty for analysis. There are only a few datasets which have geocoded violent event data and between them there exist significant disparities in coding procedure. The events are already limited by what is reported, and from this subset, researcher-led projects versus automated conflict coding present significantly varying datasets. Machine learning based automated conflict coding tends to overreport conflict events by nature of their indiscriminate models. Researcher-led projects, while precise, are labor intensive and hence more likely to underreport conflict. This problem is often resolved by cross validation between conflict datasets [22]. However, for Kurdish Conflict data this was not possible. Between the two researcher-led databases (ACLED & UCDP-GED), only UCDP-GED provides sufficient information to filter for specifically Kurdish Conflict. This means the number of conflict events in my analysis is likely systematically under representing events with poor reporting. Hopefully, with the continued advancements in artificial intelligence, automated algorithms better identify conflict events, making their use more reliable.

Additional data constraints arise when using PRIO-GRID. Although most variables offer broad temporal coverage, their differing temporal resolutions often mean that, when multiple variables are used together, the overlapping time frame available for analysis becomes quite limited. Even when temporal coverage is large, much of the data consists of estimates measured every five or ten years, interpolated accordingly. This inevitably limits data variability, while also capturing little of the local dynamics which drive conflict.

Local spatial estimates also pose challenges for spatial regression models. The machine learning algorithms used to produce local estimates for different variables across the globe rely on a form of spatial interpolation which mimics spatial dependence in the covariates. Though this dependence violates the independence assumptions of OLS, modeling this dependence with a Durbin or SLX model for instance would measure a result of the data generating process, not the phenomenon itself. This limitation is only amplified by holes in the geographic coverage of certain variables. While I estimate these values using a common spatial interpolation approach, this tradeoff only further exacerbates covariate dependence.

As with all spatial analyses, the Modifiable Areal Unit Problem (MAUP) also must be taken into account [21]. With gridded analyses, where the resolution of aggregation is exogenous to the data, there still exists the possibility that the results are an artifice of the resolution. This may be partially resolved by aggregating groups of grid squares or interpolating within grid squares. Yet, the resolution remains contingent on the original boundaries and alternative grid configurations remain scarce.

Moving away from problems inherent to the data, there are many other tradeoffs which complicate analysis. Primarily, navigating broad geographic and temporal boundaries is particularly difficult. While papers such as Doring 2024, use a spatial panel model, data which covers multiple countries over multiple decades is often not suitable for this design. Temporal heterogeneity means accounting for a time lag may already be difficult and it is reasonable to assume that the effects of time may also be spatially heterogeneous. This may be partially remedied by creating regime based spatial panel models, but this is yet to be attempted and interpretability remains difficult. An alternative, as used in this paper, is to isolate certain time periods, a method which may even be expanded into a Spatial Differences and Differences approach. Though these periods are based around significant political events, they remain arbitrary. Further, if the analysis expects to account for spatial lag effects, it is necessary that these time periods remain relatively small. Conflict igniting neighboring conflict is a phenomenon which likely spans only a few years. Hence, in order to meaningfully account for spatial lag, a large number of observations may be reduced.

As was encountered with this analysis, the presence of too few observations ends up leading to significant difficulties. Coupled with issues such as low data variability, these issues produced a model that was inherently difficult to use for spatial analysis. What was intended to help account for meaningful lag was one of many factors which stymied any use of a spatial regression. Over the course of my analyses, I was consistently unable to run spatial specifications for my exogenous and endogenous regimes. Furthermore, I was unable to run GWR. Spatial models require their non-spatial counterparts to already capture much of the variability in the data. This means these models should not be very sparse, should have sufficient variance in their covariates, should have a sufficient quantity of data, and generally should have a moderately high R squared. Without meeting these requirements, the matrices required to perform spatial analysis are often non-invertible, making this analysis impossible.

However, there is indeed a workaround to this methodological limitation. Because individual regimes derived from my spatial constrained endogenous regimes analysis had strong enough models, I was able to ex post fit spatial models to the data derived from OLS based regime creation. These specifications found that one to two of these models would benefit from a spatial specification. While one of these models was largely inconclusive and included the aforementioned suspicious covariate lag, the other presented a strong indication of a spatial spillover effect and its inclusion greatly strengthened the model. This only underscores the necessity of analysis for spatial heterogeneity as a means of fitting meaningful spatial regression models. This being said, fitting a spatial model to an OLS derived endogenous regime is not quite as meaningful as producing the endogenous regime with a spatial model. As the generative process is not tied to the spatial model, the division of the regimes is not based on any spatial specification. While this presents a significant methodological drawback, this alternative nevertheless provides meaningful information about the spatial dynamics in the data.

Based on the analysis in this paper, I offer a series of suggestions for future researchers when grappling with these methods: Firstly, it might be ideal to focus on populations for which multiple conflict datasets can be cross compared. Cross-comparison coupled with the potential use of models meant to account for zero inflated data should create units of observations with greater variance in the outcome variable. Secondly, it is important to make sure there is a sufficient amount of variation for the covariates. While there still exists limited data at such a small resolution, researchers should take great account of what available options exist and prioritize that data which has greater geographic and temporal variance. On this note, data researchers should be very cautious of fitting any model which accounts for covariate based dependence. If this dependence is derived from data generation which utilizes machine learning, these dependencies are likely fabrications of the data generating process. Finally, it is often important to examine time periods in which there is no reason to believe the data sufferers from both spatial and temporal heterogeneity. While each individually may be accounted for, accounting for both may require a Spatial Difference and Difference approach, which is unsuitable for many research questions.

CHAPTER 6

CONCLUSION

As an increasing number of datasets are created which provide sub-national gridded variables, it is important to consider how this data should be applied to models in Conflict Studies. There exist theoretical reasons to believe that spatial econometric models are well suited to model both autoregression and the diverse set of dependencies which exist between this spatially structured data, yet these applications are few and far between in the literature.

This paper discusses a primary cause for this under representation – spatial heterogeneity. Spatial heterogeneity, which has yet to be incorporated into Conflict Studies modeling, yields great benefits in its illumination of spatial dynamics and its importance in implementing spatially expansive models. Owing to the nullification and interpretability issues rendered by spatial heterogeneity, neglecting to account for non-stationarity leaves many potentially novel results unrealized. This paper provides an overview of the types of spatial heterogeneity, continuous and discrete, and the most updated toolbox of methods which exist to incorporate analysis of nonstationarity.

By analyzing the Kurdish Conflict from 1980-2015, the essential nature of spatial heterogeneity techniques is affirmed. The Kurdish conflict data exhibits strong spatial effects, but is largely meaningless before the incorporation of a spatial constrained endogenous regimes analysis. However, the ability to look at both spatial dependence coupled with heterogeneity is stymied by difficulties with the data and the model. Primarily, because the conflict data is scarce, the data has low variance, and the time constraints needed for a meaningful spatial lag render the model with few observations. This limitation is partially mitigated by the use of an alternative method, wherein an ex-post spatial specification is applied to individual OLS derived endogenous regimes. While this analysis provides meaningful results, demonstrating that one of the regimes greatly benefits from the addition of a spatial lag effect, this approach is nevertheless less robust than if the endogenous regimes were run using a spatial model.

Issues such as these are addressed at the end of the paper by a series of recommendations to future researchers on how to more successfully analyze sub-national conflict data while accounting for spatial effects. Moving forward, cross comparison of datasets, methods which account for zero inflation, caution surrounding variation and dependence in covariate data, and a general avoidance of questions with data containing both spatial and temporal heterogeneity, may prove useful in the more successful account of spatial effects in sub-national conflict studies work.

REFERENCES

[1] Martin Andresen. Theorizing globally, but analyzing locally: the importance of geographically weighted regression in crime analysis. *Crime Science*, 11(10), 2022.

[2] Luc Anselin. *Spatial Econometrics: Methods and Models*. Springer Dordrecht, 1988.

[3] Luc Anselin. Thirty years of spatial econometrics. *Papers in Regional Science*, 89(1):3–26, 2010.

[4] Luc Anselin and Pedro Amaral. Endogenous spatial regimes. *Journal of Geographical Systems*, 26:209–234, 2024.

[5] Ofra Bengio. Anfal and Halabja Genocide: Lessons Not Learned. *The Journal of the Middle East and Africa*, 14(3):277–298, September 2023.

[6] Hamit Bozarslan. *The Cambridge History of The Kurds*. Cambridge University Press, 2021.

[7] Halvard Buhaug. Local determinants of African civil wars, 1970e2001. *Political Geography*, 25:315 – 335, 2006.

[8] Stefan Doring. Spatial patterns of communal violence in sub-Saharan Africa. *Journal of Peace Research*, 61(5):858–873, 2023.

[9] J. Paul Elhorst. *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. Springer Berlin, 2014.

[10] A S Fotheringham. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Environment and Planning*, 28(4):281–298, 1996.

[11] A S Fotheringham. Multiscale Geographically Weighted Regression (MGWR). *Annals of the American Association of Geographers*, 107(6):1247–1265, August 2017.

[12] Hao Guo. A generalized regionalization framework for geographical modelling and its application in spatial regression. *arXiv*, June 2022.

[13] Mariaflavia Harari. Conflict, Climate, and Cells: A Disaggregated Analysis. *The Review of Economics and Statistics*, 100(4):594–608, October 2018.

[14] Havard Hegre. Population Size, Concentration, and Civil War: A Geographically Disaggregated Analysis. *World Bank*, (4243), 2007.

[15] Håvard Hegre. Poverty and Civil War Events A Disaggregated Study of Liberia. *Journal of Conflict Resolution*, 53(4):598–623, August 2009.

[16] Marco Helbich. Data-Driven Regionalization of Housing Markets. *Annals of the American Association of Geographers*, 103(4), July 2013.

[17] Cullen Hendrix. Climate change, rainfall, and social conflict in Africa. *Journal of Peace Research*, 49(1):35–50, 2012.

[18] Magnus Theisen Ole. Climate clashes? Weather variability, land pressure, and organized violence in Kenya, 1989–2004. *Journal of Peace Research*, 49(1):81–96, 2012.

[19] John O'Loughlin. A Review and Assessment of Spatial Analysis and Conflict: The Geography of War, 2009.

[20] John O'Loughlin. The diffusion of violence in the North Caucasus of Russia, 1999–2010. *Environment and Planning*, 44:2379–2396, 2012.

[21] Stan Openshaw. The Modifiable Areal Unit Problem. In *CONCEPTS AND TECHNIQUES IN MODERN GEOGRAPHY*, volume 38. Geo Books, 1983.

[22] C Raleigh. Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices. *Humanit Soc Sci Commun*, 2023.

[23] Clionadh Raleigh. Introducing ACLED: An Armed Conflict Location and Event Dataset. *Journal of Peace Research*, 47(5), 2010.

[24] Clionadh Raleigh and Håvard Hegre. Population size, concentration, and civil war. A geographically disaggregated analysis. *Political Geography*, 28:224–238, 2009.

[25] Guy Schvitz. Mapping The International System, 1886-2017: The CShapes 2.0 Dataset. *Journal of Conflict Resolution*, 66(1):144–61, 2022.

[26] Ralph Sundberg. Introducing the UCDP Georeferenced Event Dataset. *Journal of Peace Research*, 50(4), 2013.

[27] W.R Tobler. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46:234–240, June 1970.

[28] Andreas Forø Tollefsen. PRIO-GRID: A unified spatial data structure. *Journal of Peace Research*, 49(2):363–374, 2012.

[29] Andreas Forø Tollefsen. PRIO-GRID v.2.0 codebook, 2015.

[30] Michael Townsley. Space Time Dynamics of Insurgent Activity in Iraq. *Security Journal*, 21:139–146, 2008.

[31] Wladimir van Wilgenburg. From the PYD-YPG to the SDF: the Consolidation of Power in Kurdish-Controlled Northeast Syria. *Studies in Conflict & Terrorism*, 47(9):1090–1109, 2024.

[32] Nils B. Weidmann. Diffusion patterns of violence in civil wars. *Political Geography*, 30:143–152, 2011.

[33] Julian Wucherpfennig. Politically Relevant Ethnic Groups across Space and Time: Introducing the GeoEPR Dataset. *Conflict Management and Peace Science*, 28(5):423–37, 2011.

APPENDIX A

VARIABLE	DF	VALUE	PROB
CONSTANT	4	10.913	.0276
log_urban_ih_filled	4	3.543	.4714
petroleum_y	4	1.681	.7941
bdist3	4	20.145	.0005
capdist	4	25.599	.0000
Global text	**20**	**84.727**	**0.0000**

Table A.1: Chow Test Results for 1997-2003 Regime Diagnostics

Variable	Coefficient	Probability
CONSTANT	-.00133	.0001
log_urban_ih_filled	.00047	.00052
petroleum_y	.00474	.0000
bdist3	.00000	.03837
capdist	.00000	.00005
W_log_k_rate	**.30317**	**.01839**
Spatial Pseudo R-squared	.4873	

Table A.2: Spatial Lag Model applied to Regime 2 (2004-2009) from the original Endogenous Regimes Process