Anna Glass
DATA 150 Evolving Solutions
Professor Frazier
13 November 2020
Word Count: 2,767

## Introduction

Across the world, failing governments systemically deny civil rights and political liberties to citizens. Freedom and representation exist in an inextricable entanglement; states cannot improve a population's well-being without an accurate measurement of where the population suffers.[1] Migratory census data in particular serves to quantify a location's propulsiveness and providing valuable insight into deterring and attracting features of communities. Data scientists have turned to spatially disaggregated population estimates in areas where geospatial data remains unreliable or infrequent. States struck with war and poverty, particularly those in the African and Middle Eastern regions, suffer from a critical lack of information. In Lebanon, the most recently enumerated census data comes from 1932. In Somalia, the only public census occurred in 1975.[2] Political instability and civil unrest preclude census takers from performing their research, and corrupt governments may block procured data. This instability and its connections to population and demographic shifts would elucidate quantitative indicators of human and political well-being. When domestic instability inhibits the procuring and implementation of census-driven conclusions, government-collected macro-census data is no longer sustainable.

Synthetic population data provides a realistic opportunity for researchers to obtain data concerning migratory patterns. Census microdata provides high-resolution, anonymized samples of national census data and allows for a more robust migration model in underrepresented regions. This analysis will explore two methods: the gravity-type spatial interaction model (GTSIM) and hierarchical Bayesian modeling for uncertainty. The gravity-type spatial interaction model can track human movement using sparse census data, while hierarchical Bayesian modeling can identify and adapt to irregularities and uncertainty in migratory census data and create a confidence interval that better informs organizational response.[3] This analysis will expand the central research question: *Within the context of population shifts in the Middle East and Africa, what is the impact of political crises such as terrorism, government instability, or climate change?*

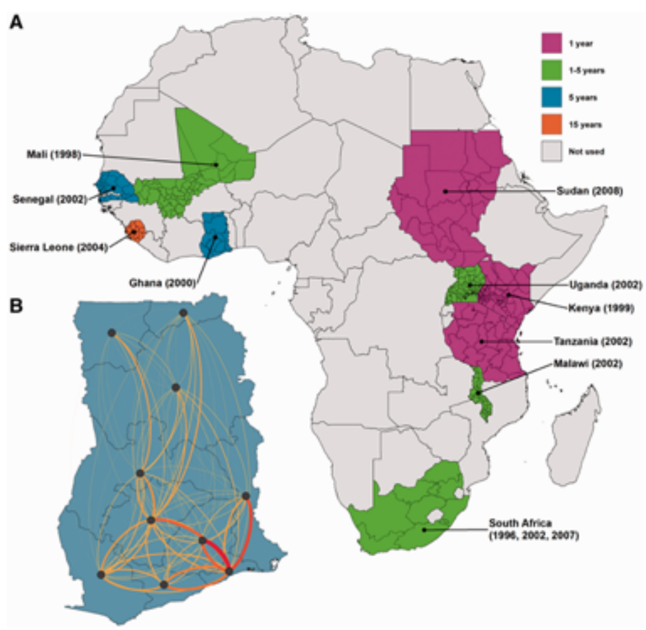
## Support for Selected Type of Inquiry

The Middle East and African regions face a critical lack of procured census data, which clouds essential patterns that can explain and predict the relationship between external factors and human movement. To address this gap in information, I propose a three-pronged central research question that concerns itself with: 1) how political crises impact migratory behavior in Africa and the Middle East; 2) the predictive nature of migratory patterns in relation to political crises, and; 3) the necessary considerations to account for uncertainty unenumerated populations. This central research question trends toward an explanatory inquiry; it seeks to identify causes and effects to understand patterns in migratory behavior. In particular, this inquiry aims to explain the processes that determine a location's propulsiveness to understand systematic patterns in migratory behavior. Causal relationships in human migration bring light to domestic and international push-pull factors, specifically the relationship between observed political events and migratory data.

This central research question expands further to address its inverse: How does power transition theory explain historical migratory patterns, and how can it predict future demographic shifts? The sub-research question yields itself to a causal or predictive puzzle inquiry; it explores the influence of particular factors on larger themes and seeks to explain the observed phenomenon of human migration with a multilateral approach. The second sub-research question aids the broader research question: How can current migratory and climate data predict the long-term movements that the Middle East and Africa will see as a result of global warming, and how might this give rise to political and economic conflict? The third sub-research question serves as an investigation into a gap in the current methods: What supplementary methods can be used in combination with CDRs and spatial interaction models to account for unrepresented populations? A fourth sub-research question addresses these findings' future applications: What value do the findings surrounding political crisis event-related migration have for developing sustainable communities? The gravity-type spatial interaction model and hierarchical Bayesian modeling framework that accounts for uncertainty combine to address each question.

## The Gravity-Type Spatial Interaction Model

The gravity-type spatial interaction model (GTSIM) can track human movement using sparse census data. Coherent with large-scale spatially disaggregated population datasets, the GTSIM incorporates datasets from population sizes at each microcensus location.[4] The model uses the distance between population concentrations to evaluate which location characteristics contribute to within-country migration, distancing itself from previous methods that correlated population with a location's propulsiveness. Two principle studies evaluate migration flows using the gravity-type spatial interaction model.




### Andres Garcia

First, Andres Garcia and his team aim to understand the driving factors behind migration in Sub-Saharan Africa by modeling internal migration flows in sub-regions that remain poorly quantified—national migratory data is more widely available but does not record the nuances of within-region migratory patterns.[5] Garcia obtained census microdata from Integrated Public Use Microdata Series (IPUMS) from locations in sub-Saharan Africa. Census participants changed residency across striking variations in spatial and temporal resolution. Garcia queried the census microdata using Structured Query Language to quantify migrations between administrative units i and j. The model measures propulsiveness using other demographic and socioeconomic variables with established relations to migration— namely precipitation, calculated with a 0.5-degree spatial resolution, modeled drought and rainfall variability across regions. The GTSIM operates based on spatial interaction theory's postulation that, concerning migration, individuals seek to minimize costs and maximize benefits. Thus, the dependent variable $MIG_{ij}$ observes the flux in migration flows between two administrative units with the assumption that $MIG_{ij}$ is proportional to the locations' masses and inversely proportional to the physical distance between them. This presents the equation:

$$MIG_{ij} = \frac{p_i^{\alpha} p_j^{\beta}}{d_{ij}^{\gamma}}$$

Where the variables $p_i$ and $p_j$ represent the population size at an origin *i* and a destination *j*. The variable $d_{ij}$ refers to the distance between points i and j, while the exponents $\alpha$, $\beta$, and $\gamma$ are obtained from the data and unknown.[6]

**Amy Wesolowski**

In cooperation with Garcia, Amy Wesolowski delves into the application of census migration data to estimate human movement patterns across temporal scales. Improved census data using spatial interaction models can reveal shorter-scale migration patterns, including seasonal relocations or movements lasting less than one year.[7] Although national census data acts as a proxy for shorter-term movements, mobile phone usage data expands this highly sensitive dataset's analytical breadth. In addition to quantifying general regional movement, call data records (CDRs) construct an empirical density distribution using the Euclidean distance between countries to quantify their ranking as source countries and destination countries. Wesolowski's gravity model, can be described by the equation:

$$N_{x,y} = \frac{population_x^{\alpha} * population_y^{\beta}}{dist(x,y)^{\gamma}} + k$$

Where x and y represent two distinct locations, and $population_x$ *and* $population_y$ represent their respective locations' populations. The exponents $\alpha, \beta, \gamma$ and the intercept k were included in the equation after fitting the model to verified data using a generalized linear model with a Poisson specification. The distance exponent $\gamma$ and the destination's population exponent $\beta$ serve to model the relationship between the duration of the journey and the relative importance of the destination.[8] The gravity model's exponents successfully approximate a variety of movements across multiple temporal scales with only enumerated national census data. In this model, the two factors that estimate movement are physical distance measured with Euclidean coordinates and importance, represented by each location's population size.

The model in Wesolowski's study includes both a spatial and temporal component, with comprehensive comparisons of inter-region trips in Kenya, derived across multiple temporal scales over the course of a year. The absolute number of trips were compared to census data using the Pearson's correlation coefficient and linear regression, with the ultimate product of an empirical density distribution based on the Euclidean distance between a pair of counties. The study concluded that the model most closely predicted movement in trips between three and four months, and total movement with adjusted $R^2$ values ranging from 0.134 to 0.404. The linear fits for country rankings revealed adjusted $R^2$ values between 0.45–0.67 and 0.32–0.60.[9] The flow ranks between counties revealed a closer fit than absolute movement values. Despite expected inaccuracies in movements over complex timescales, the ordering of counties by movement reveals a strikingly similar dataset to national census collection. These findings create an imperative for researchers to evaluate further the relationship between sociodemographic variables, climatic variables, and migratory patterns.

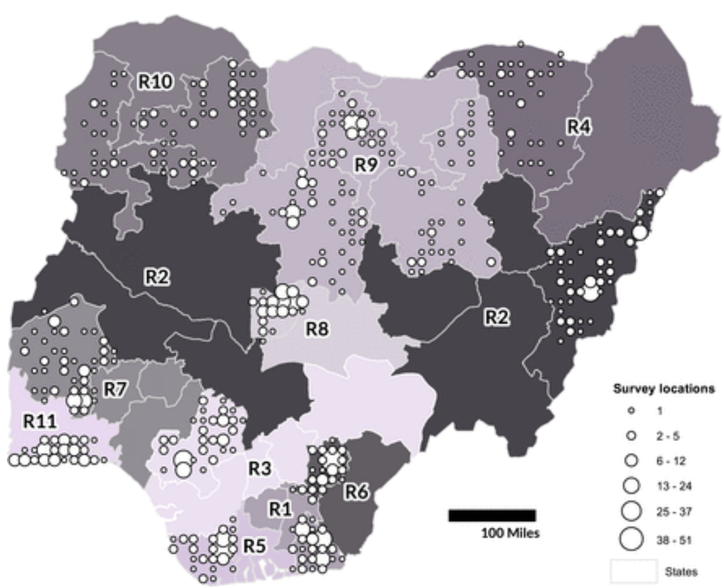
## Hierarchical Bayesian Modeling

In order to adopt gridded population survey data in human development policy, the methods must feasibly be able to be implemented across diverse settings and account for cell-level accuracy.[10] Emerging Bayesian modeling methods aim to harmonize an archive of geospatial datasets across large spatial areal units by representing the population as a continuous surface instead of arbitrary administrative boundaries.[11] Hierarchical Bayesian modeling complements the gravity-type spatial interaction model, emphasizing quantifying uncertainties and creating comparative confidence intervals for estimated population data. A foundation of microcensus data, simulated annealing, and analysis of CDRs are necessary to understand and further population-focused Sustainable Development Goals (SDGs).

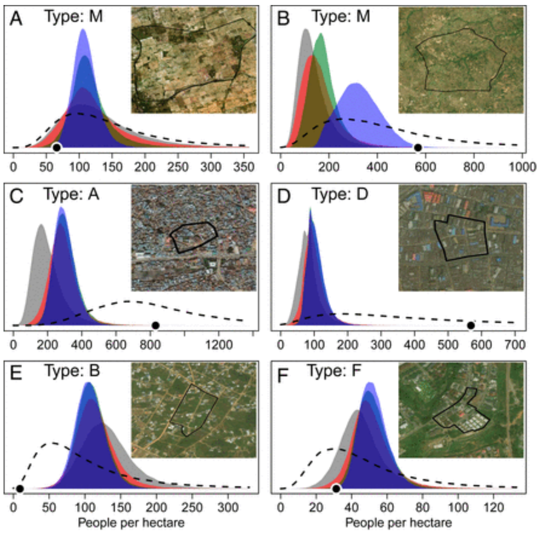
Simulated population data and alternative census collection methods reveal a panacea for representation in demography. Missteps in the domain of preserving human life can prove fatal. Thus, the confounding variable of uncertainty proves the largest hindrance to the full implementation of national population mapping using sparse survey data. Hierarchical Bayesian modeling framework accounts for uncertainty in predictive population mapping, which progresses the proposed methods from theory to execution.[12] The hierarchical Bayesian model combines neighborhood-scale microcensus surveys in Nigeria with national-scale satellite imagery.

### Douglas Leasure

Douglas Leasure's study combines the United Nations Population Division's annual projection models' development goals and disaggregate population totals realized through satellite imagery and geospatial data to develop uncertainty intervals for past population estimates. Bayesian statistics can customize models for microcensus and other population data to account for complex relationships such as non-Gaussian error distribution—a probability distribution of random error that does not follow a normal distribution—or random effects and observer error.[13]




The Bayesian modeling framework combines geospatial covariates with data from recent microcensus surveys to predict population sizes in Nigeria's unobserved areas. Geospatial




covariates include local government areas (l), states (s), and regions (r), and accounted for subjective definitions of a region in collaboration with local experts. Population estimates contained average error rates from 67 to 92 people per hectare for population data predicted out-of-sample. A posterior probability distribution for random intercepts $\alpha_{t,r,s,l}$ reveal density for six types of settlement types within a Nigerian administrative unit. Blue represents densities for a settlement type, red represents microcensus survey population density, and the dashed like Di represents population density estimates.[14] In its most basic form, Leasure measures population density by extending the Poisson process model, $N_i \sim Poisson(D_i A_i)$, which models counts of people, to $D_i \sim LogNormal\left(\bar{D}_i, \sigma_{t,r,s,l}\right)$. Administrative unit types A B D and F account for distinct urban areas an type M accounts for rural areas.[15] The hierarchical Bayesian modeling for uncertainty creates population density estimates on a 100-m spatial resolution grid.

**Findings**

Concerning reducing uncertainty, three geospatial covariates had significant positive effects: WorldPop ($x_1$) gridded population estimates, school densities ($x_2$), and household sizes ($x_3$). Conversely, total settled area ($x_4$), residential area ($x_5$), and nonresidential settlement ($x_6$) had insignificant effects on the confidence interval. Leasure found that estimates of total population size yielded a lower accuracy ($r^2$ of 0.46) than estimates of population density ($r^2$ of 0.26), which he attributes to outdated estimates of settled area. The model overestimates slightly with a positive bias in areas with low populations.

Despite this imprecision, the model proves useful; in cross-validated model predictions using comparably data-sparse areas at a spatial scale of 3 ha, the average imprecision was 121 people per hectare and overall error was 92 people per hectare. The hierarchical modeling framework's greatest strength is its acknowledgement of variation in population densities across differing socioeconomic contexts in Nigeria. Census data contains massive inaccuracies and outdated data that remain largely unquantified or acknowledged, and hierarchical Bayesian modeling promises to harness the uncertainty to guide high and low social planning scenarios.[16] Uncertainty varies along with population, and patterns in uncertainty yield a guide to preparedness across multiple levels of precision. Hierarchical Bayesian modeling can coalesce with additional census data to support a "living census" through time series models.

## Development of Central Research Question

The GTSIM and hierarchical Bayesian modeling have evident and countless applications in predicting and supplementing demographic and migratory data in underrepresented areas. Although my previous inquiry investigated synthetic population data in relation to tracking migratory patterns and the spread of disease, I now aim to further quantify migratory patterns and their causal relationship with political events. Developments in mass data would suggest that data science can quantify and predict relationships between seemingly abstract concepts—in this case, subcategories of political events and the reaction to climate change. The causes and conditions of international conflict and war have been long studied, but rarely quantified with scientific consensus. By combining the GTSIM and hierarchical Bayesian modeling, I aim to identify patterns in migration unique to three subcategories of political events—inter/intra-state war, terrorist insurgency, and disaster migration—labelling each event according to general historical consensus.

The GTSIM serves to answer the central research questions because of its predictive and causal capabilities. Garcia's findings have primary applications in quantifying and modeling human migration better understand policy, economic development, and communicable disease. Garcia demonstrates the GTSIM's applications in quantifying a population subset's social and demographic landscape, which indicates a clear step toward quantifying a political landscape. The GTSIM can explain up to 87.4% of migration after successful deviance reduction and iteration.[17] Further journals have used the GTSIM to model transnational terrorism's consequences on trade, which creates a precedent to quantify terrorist perpetrators and their relation to intra- and inter-country migratory patterns.[18] Historically, F. Sà used the GTSIM to measure the relationship between trade and domestic factors such as ethnic heterogeneity and international economic engagement—each factors that determine political stability.[19] The GTSIM shows explanatory and predictive promise in quantifying political crises with the incorporation of further conflict and power transition theory.

Further, hierarchical Bayesian modeling has a longstanding relationship with comparative politics and cross-sectional time-series models. It has historically combined theories of comparative political economy to explain causal complexity in political processes. Bayesian hierarchical models obtain optimal statistical estimates to evaluate how institutional conditions interpret demographic, social, and environmental variables. Hierarchical Bayesian modeling interprets a variable's contextual effects, which accounts for uncertainty and rare political events such as terrorist insurgency. Garrett Alvarez illustrated the hierarchical model's applications in political determinant reanalysis of economic growth in OECD countries; the model has historical merit in the Middle Eastern region and could extend its statistical capabilities to explain and predict political crises.[20] Hierarchical Bayesian modeling validates itself with its emphasis on accounting for uncertainty. The model makes assumptions about populations and factors those assumptions into conclusions to present a reliable confidence interval. Probabilistic estimates of uncertainty reduce false confidence in often unreliable census data; the GTSIM and Bayesian hierarchical model combine to create high-resolution, cross-tabulate data with an awareness of uncertainty and irregularity.

## Identification of a Plausible Research Gap

One thousand people per hour fled from Libya to Egypt and Tunisia after the Arab Spring; its effects—two million displaced people—continue to ripple across North Africa and the Middle East.[21] Conflict has tangible impacts on migration and displacement; however, these impacts often remain overlooked in analyzing major international events. Data science in its breadth and scope of migratory data has the potential to determine the nature of an uprising and expedite intervention from international bodies. The largest obstacle in obtaining this goal? Synthetic population data lacks an understanding of the theoretical literature that drives migratory patterns. Specifically, the field would benefit from synthesizing demographic and international relations theory with the procured data that indicates migratory push-pull factors. Microcensus data combined with CDRs widens the scope and frequency of migration over time and allows for distance comparison. Though these models seemingly capture a habitual, accurate delineation of population patterns, both methods contain crude elements.

Garcia's employment of the gravity-type spatial interaction models represents a step forward but requires severe contextualization. The push-pull factors within a region remain heavily individualized and difficult to quantify on a large scale without implementing survey data alongside mapping. Consider that improved conditions, typically associated with industrialization, lead to higher costs of living. In this case, migratory "push" factors may indicate improved HDIs despite the appearance that the location deters residents. Similarly, Wesolowski's study requires verification that cell phone data usage does not introduce confounding variables. Individuals that regularly engage with cell phone towers represent a wealthier subsection of the population in areas where mobile phone usage remains uncommon. The field has slowly ameliorated its incomplete understanding of small-scale migratory patterns by implementing methods across high temporal and spatial resolutions. To model the relationship between political movements and population movements, migratory theory and mass data must coalesce to identify common determinants of revolt and emigration, and whether these phenomena have roots in demographic change.[22]

## References


[1] Sen, A. (2019). *Development as freedom*. MTM.

[2] Wardrop, N. A., Jochem, W. C., Bird, T. J., Chamberlain, H. R., Clarke, D., Kerr, D., … Tatem, A. J. (2018, April 3). *Spatially disaggregated population estimates in the absence of national population and housing census data*. PNAS.

[3]Anderson J E. "The Gravity Model." SSRN eLibrary, 2010 ; Leasure, D. R., Jochem, W. C., Weber, E. M., Seaman, V., & Tatem, A. J. (2020, September 29). *National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty*. PNAS. https://www.pnas.org/content/117/39/24173.

[4] Anderson J E. "The Gravity Model." SSRN eLibrary, 2010

[5] Garcia, A. J., Pindolia, D. K., Lopiano, K. K., & Tatem, A. J. (2014, August 4). *Modeling internal migration flows in sub-Saharan Africa using census microdata*. Migration Studies. https://academic.oup.com/migration/article/3/1/89/2413406.

[6] Garcia, A. J., Pindolia, D. K., Lopiano, K. K., & Tatem, A. J. (2014, August 4). *Modeling internal migration flows in sub-Saharan Africa using census microdata*. Migration Studies. https://academic.oup.com/migration/article/3/1/89/2413406.

[7] Wesolowski, A., Buckee, C. O., Pindolia, D. K., Eagle, N., Smith, D. L., Garcia, A. J., & Tatem, A. J. (2013, January 9). *The Use of Census Migration Data to Approximate Human Movement Patterns across Temporal Scales*. PLOS ONE.

[8] Wesolowski, A., Buckee, C. O., Pindolia, D. K., Eagle, N., Smith, D. L., Garcia, A. J., & Tatem, A. J. (2013, January 9). *The Use of Census Migration Data to Approximate Human Movement Patterns across Temporal Scales*. PLOS ONE.

[9] Wesolowski, A., Buckee, C. O., Pindolia, D. K., Eagle, N., Smith, D. L., Garcia, A. J., & Tatem, A. J. (2013, January 9). *The Use of Census Migration Data to Approximate Human Movement Patterns across Temporal Scales*. PLOS ONE.

[10] Thomson, D. R., Rhoda, D. A., Tatem, A. J., & Castro, M. C. (2020, September 9). *Gridded population survey sampling: a systematic scoping review of the field and strategic research agenda*. International Journal of Health Geographics. https://ij-healthgeographics.biomedcentral.com/articles/10.1186/s12942-020-00230-4.
[11] Lloyd, C. T., Chamberlain, H., Kerr, D., Yetman, G., Pistolesi, L., Stevens, F. R., … Tatem, A. J. (2019, June 18). *Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets*. Taylor & Francis. https://www.tandfonline.com/doi/full/10.1080/20964471.2019.1625151.
[12] Stewart, Robert, et al. "A Bayesian machine learning model for estimating building occupancy from open source data." *Natural Hazards* 81.3 (2016): 1929-1956.
[13] Leasure, D. R., Jochem, W. C., Weber, E. M., Seaman, V., & Tatem, A. J. (2020, September 29). *National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty*. PNAS. https://www.pnas.org/content/117/39/24173.
[14] Bryant, John R., and Patrick J. Graham. "Bayesian Demographic Accounts: Subnational Population Estimation Using Multiple Data Sources." *Bayesian Analysis*, International Society for Bayesian Analysis, 2013, projecteuclid.org/euclid.ba/1378729921.
[15] Leasure, D. R., Jochem, W. C., Weber, E. M., Seaman, V., & Tatem, A. J. (2020, September 29). *National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty*. PNAS. https://www.pnas.org/content/117/39/24173.
[16] Leasure, D. R., Jochem, W. C., Weber, E. M., Seaman, V., & Tatem, A. J. (2020, September 29). *National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty*. PNAS. https://www.pnas.org/content/117/39/24173.
[17] Garcia, A. J., Pindolia, D. K., Lopiano, K. K., & Tatem, A. J. (2014, August 4). *Modeling internal migration flows in sub-Saharan Africa using census microdata*. Migration Studies. https://academic.oup.com/migration/article/3/1/89/2413406.
[18] Bandyopadhyay, Subhayu, et al. "Trade and Terrorism: A Disaggregated Approach - Subhayu Bandyopadhyay, Todd Sandler, Javed Younas, 2018." *SAGE Journals*, 17 Apr. 2018, journals.sagepub.com/doi/full/10.1177/0022343318763009.
[19] Sà, F. *Gravity Model*, ScienceDirect, 2013, www.sciencedirect.com/topics/economics-econometrics-and-finance/gravity-model.
[20] Western, Bruce. "Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach." *American Journal of Political Science*, vol. 42, no. 4, Oct. 1998, pp. 1233–1259., doi:10.2307/2991856.
[21] Koser, Khalid. "Migration, Displacement and the Arab Spring: Lessons to Learn." *Brookings*, Brookings, 28 July 2016, www.brookings.edu/opinions/migration-displacement-and-the-arab-spring-lessons-to-learn/.
[22] Fargues, Philippe. "Mass Migration and Uprisings in Arab Countries: An Analytical Frame..." *International Development Policy | Revue Internationale De Politique De Développement*, Institut De Hautes Études Internationales Et Du Développement, 1 Jan. 2017, journals.openedition.org/poldev/2275.