"Spatiotemporal Modeling of Economic Activity in Cuba from Satellite Proxies": Stat 288 Midterm

Christopher Perez

March 25, 2025

1 Short Essay

Measuring economic well-being in Cuba is notoriously difficult due to scarce and unreliable official statistics. The World Bank's open data portal lacks many poverty and inequality metrics for Cuba, and even basic metrics like GDP growth are contested—recent figures suggest an annual growth rate of just 1.4% (World Bank 2024). In the absence of trustworthy ground data, researchers increasingly turn to satellite-based proxies. Nighttime lights, for example, strongly correlate with economic output in data-sparse regions, and Cuba's contrast to Florida in nighttime imagery highlights its economic constraints. Yet Cuba presents a uniquely different case: government data is limited and politically controlled, independent surveys are rare, and access to the island's interior remains logistically and digitally restricted. These limitations make traditional poverty measurements infeasible and motivate the use of remote, indirect methods. *How can we infer latent economic activity across Cuba using multi-source geospatial data in a setting with no reliable ground truth?*

I propose a multi-source and Bayesian spatial modeling approach to estimate a latent "economic activity" index across the island. Three remote proxies will be used: (1) nighttime lights, reflecting urbanization; (2) vegetation indices, capturing agricultural productivity; and (3) OpenStreetMap infrastructure data, such as road density, which have been linked to city-level GDP (Liu et al. 2020).

The main statistical task is to infer a spatiotemporal latent economic activity index, denoted by z_{it}, for each spatial location $i \in \{1, \ldots, N\}$ and time point $t \in \{1, \ldots, T\}$, using satellite-derived proxies. I will develop a Bayesian hierarchical model in which each proxy x_{ijt} (e.g., nightlights, road density) is modeled as a function of the latent economic variable z_{it}, such as

$$x_{ijt} \mid z_{it}, \beta_j, \sigma_j^2 \sim \mathcal{N}(\beta_j z_{it}, \sigma_j^2),$$

where β_j captures the strength of association between proxy j and the latent economic signal, and σ_j^2 accounts for error.

To model the latent structure, I will also impose spatial and temporal dependence. For spatial regularization, the latent variables $\mathbf{z}_t = (z_{1t}, \ldots, z_{Nt})$ at each time t will follow a Gaussian Markov Random Field prior, so that nearby locations have similar economic ac-

tivity levels (Steele et al. 2017). Temporally, I will model z_{it} as evolving smoothly over time using an autoregressive prior.

Information from empirical studies in comparable low-income countries will inform the distributions of the coefficients β_j. The model will be implemented in Python using PyMC or Stan, with GPUs from FAS Research Clustering. The output will be a posterior distribution over economic activity at each spatial location, from wich I will generate both point estimates and uncertainty estimates.

This project aims to produce the first data-driven map of economic activity in Cuba, with applications in spatial inequality analysis. I hope to highlight underdeveloped regions and quantify where inference is weakest. Beyond Cuba, I seek to generalize the methodology to other low-data contexts, providing a blueprint for economic estimation in data-constrained environments.

2 Project Assessment and Plan

(a) Data

I will use three publicly available datasets:

- **Nighttime Lights:** VIIRS data from NOAA (link).
- **Vegetation Indices (NDVI):** MODIS NDVI data from NASA's LP DAAC (link).
- **Infrastructure:** Road network density from OpenStreetMap via Geofabrik (link).

Given the large size of the satellite imagery, I can store the data on FAS Research Computing clusters. I can also compress the raw files and used chunked data processing to reduce memory usage issues.

Data will be spatially aligned into a grid using Python libraries.

Ethical concerns are minimal since these datasets are openly licensed and aggregated, however I will ensure compliance with data usage policies.

(b) Computation

I will perform computations on Harvard's FAS GPU clusters. Bayesian inference will use Python tools (PyMC or Stan). I have prior experience using GPU-based model training (for a prior RoBERTa model project) and experiment tracking with WandB. I will use the latter to monitor model performance and hyperparameter tuning.

(c) Plan

Over the next five weeks leading up to the May 5 presentation, I will follow this rough timeline to complete the project.

- **Week 1:** Acquire and preprocess datasets.
- **Week 2:** Conduct EDA and implement a basic Bayesian hierarchical model without temporal structure. Try adding the temporal structure at this point to see if it is or is not feasible? Based on my initial findings, I will reassess the project scope.
- **Week 3:** Research Gaussian Markov Random Fields and integrate into project for spatial correlation. Run the spatial model and evaluate computational feasibility.
- **Week 4:** If feasible, integrate temporal structure; otherwise, rigorously refine the spatial-only model. Run diagnostics and validate model results. Track results with WandB. All code will be stored in a Github repo.
- **Week 5:** Generate final visualizations and uncertainty maps. Maybe develop an interactive web visualization using React?

(d) Skills and Feasibility

This project aligns well with my Python skills, prior experience with Bayesian inference basics from previous statistics courses and a research assistantship, GPU computations, and WandB experiment tracking. Implementing GMRFs will provide additional statistical depth, but will require me doing more background research so I know how to best integrate this into my project.

Due to computational constraints, the temporal component may be simplified if needed. This will be something I will assess as soon as possible to know if I need to adjust my project scope.

(e) Challenges

Main challenges include the computational complexity of the spatiotemporal model and the statistical complexity of the GMRF implementation. If complexity proves difficult, I will simplify by using a purely spatial model or aggregated data. If data quality is poor, I will explore alternative datasets (this could very well be an issue with data from Cuba). Another challenge will be to weigh the three sources appropriately in the model, but will hopefully provide useful insights.

(f) Contribution

I seek to demonstrate how multi-source satellite proxies can be combined in a Bayesian hierarchical framework to estimate latent economic activity in the absence of reliable ground truth. I may additionally offer insights into model design, proxy weighting, and the tradeoffs involved in adding spatial and temporal structure under computational constraints.

I also hope to generate new spatial knowledge about regional economic disparities in Cuba, offering the first estimates of economic activity throughout the island. The project will highlight underdeveloped regions and areas of high uncertainty, with potential applications in future research on low-data environments!

References

Liu, B., Shi, Y., Li, D.-J., Wang, Y.-D., Fernandez, G., & Tsou, M.-H. (2020). An economic development evaluation based on the openstreetmap road network density: The case study of 85 cities in china. *ISPRS International Journal of Geo-Information*, *9*(9), 517. https://doi.org/10.3390/ijgi9090517

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.-A., Iqbal, A. M., Hadiuzzaman, K. N., Lu, X., Wetter, E., Tatem, A. J., & Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data. *Journal of the Royal Society Interface*, *14*(127), 20160690. https://doi.org/10.1098/rsif.2016.0690

World Bank. (2024). Cuba — data. Retrieved March 24, 2025, from https://data.worldbank.org/country/cuba