Inferring Cuba's Hidden Economy: A Bayesian Latent Variable Model

Christopher M. Perez

May 13, 2025

Contents

1 Introduction

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 (Steele et al. 2017).

In light of these data limitations, I adopt a Bayesian latent variable model to estimate a subnational index of economic activity in Cuba using three high-resolution proxies: nightime lights (VIIRS), normalized difference vegetation index (NDVI), and road density. This latent factor framework allows us to infer relative economic activity even without traditional survey data. Results are then cross-validated in the Dominican Republic (DR), where data is easily accessible, thereby providing a systematic method for mapping economic activity in data-poor settings.

2 Data and Preprocessing

I utilize three geospatial proxies for economic activity in Cuba, all aligned on a 500 m grid, that capture different aspects of development. All data correspond to the 2024 year.

- **Nighttime Lights (VIIRS):** Satellite-measured night light intensity, a well-known proxy for economic output. Brighter areas tend to indicate higher population density, infrastructure, and GDP.
- **NDVI (Vegetation):** The Normalized Difference Vegetation Index captures "greeness." While high NDVI often signifies productive agriculture or healthy ecosystems, it can also be inversely correlated with poverty in some contexts.
- **Road Density (OpenStreetMap):** Measures the density of roads, reflecting infrastructure. Well-connected regions tend to be more economically active.

All layers are reprojected and aligned to the same grid, and masked to exclude water or uninhabited zones. Each proxy is then standardized to account for different scales.

3 Bayesian Hierarchical Model

We define a latent variable z_i for each grid cell i, representing the underlying economic activity (on an arbitrary scale). The observed proxies are linked to z_i via a linear model described below.

3.1 Observation Model

For grid cell i and proxy k, let

$$x_{i,k} \sim \mathcal{N}\big(\beta_k \, z_i, \; \sigma_k^2\big), \qquad k \in \{\text{lights}, \; \text{NDVI}, \; \text{roads}\},$$

where β_k is a weight and σ_k is the standard deviation for proxy k. This implies that each proxy is a noisy linear indicator of the same latent z_i. For example, nighttime light intensity $x_{i,\text{lights}}$ is modeled as: $z_i \cdot \beta_{\text{lights}} + \text{noise}$.

3.2 Identifiability Constraint

To address the identifiability issue, I fix $\beta_{\text{lights}} = 1$, a unit coefficient on lights. This anchors the latent scale such that an increase of 1 in z corresponds to one standard deviation increase in night lights.

The other coefficients have weakly informative priors: $\beta_{\text{NDVI}}, \beta_{\text{roads}} \sim \mathcal{N}(0, 1)$. This allows NDVI and road density to be learned as either positive or negative indicators of z.

3.3 Spatial ICAR Prior and Neighborhood Smoothing

In our model, we include a spatial random effect modeled with an Intrinsic Conditional Auto-Regressive (ICAR) prior. This prior introduces spatial smoothing by enforcing that neighboring areas have similar random-effect values. This is achieved using the adjacency matrix of the grid cells, which is a square matrix where each element A_{ij} is 1 if cell i and cell j are neighbors and 0 otherwise. Under this prior, each area's effect is conditioned on the average of its neighbors' effects. In practice, we add a prior potential

$$\propto \; \exp\left\{ -\frac{\tau}{2} \sum_{(i,j) \in \mathcal{N}} (z_i - z_j)^2 \right\},$$

where $\mathcal{N}$ is the set of neighbor pairs, and where τ is the spatial precision parameter (Duncan et al. 2017).

The advatange of using an ICAR prior is the localized smoothing it provides. The model will produce more reliable estimates for each area by pooling information with its neighbors. This is especially helpful for areas with sparse data. Additionally, it also maintains local distinctness in that each area is mainly informed by its immediate neighbors, rather than enforcing a single global effect.

A concern with ICAR is the issue of spatial confounding, which occurs when a spatially-structured random effect competes with or masks the effects of spatially correlated covariates. In our analysis, we remain mindful of this issue so that it does not inadvertently obscure the estimated effects of our proxies.

3.4 Weakly Informative Priors

The spatial precision τ and the standard deviations σ_k parameters are modeled using weakly informative priors. Specifically $\tau, \sigma_k \sim \text{HalfNormal}(1)$. This choice allows us to have some regularization: it provides some information to prevent extreme or implausible values of the variance parameters without being too strongly constraining. This also helps improve convergence and guard against overly diffuse posteriors without contributing too much bias (Gelman 2006).

Moreover, we use a non-centered parameterization $z_i = \frac{z_i^{\text{raw}}}{\sqrt{\tau}}$, where $z_i^{\text{raw}} \sim \mathcal{N}(0,1)$, to decouple the highly correlated magnitude of z and τ.

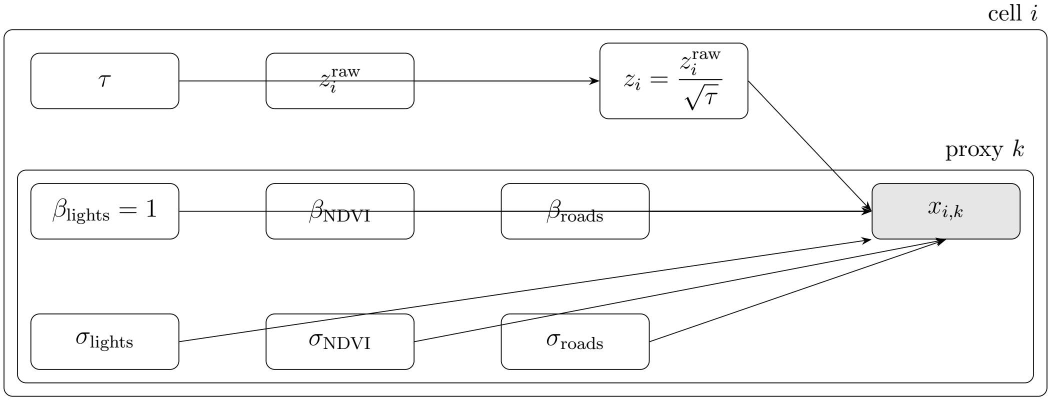
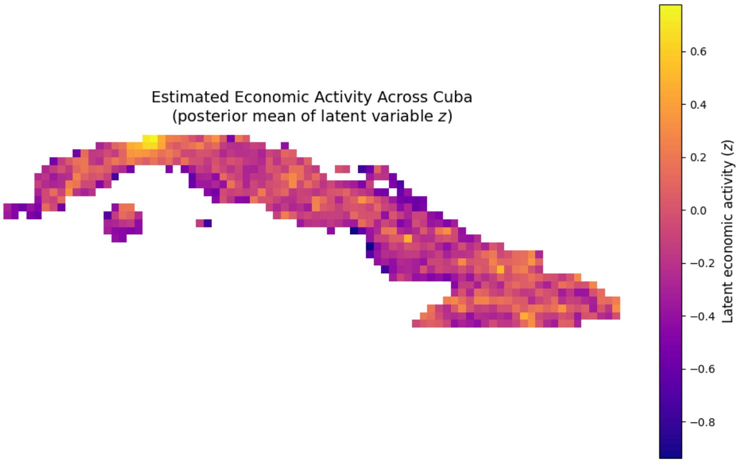


Figure 1: Graphical model of the Bayesian latent-variable hierarchy.

4 Posterior Results for Cuba (15 km Grid)

After fitting the model, we obtain the latent economic activity index z_i across Cuba at a 15 km cell resolution. Figure 2 shows the posterior mean of z_i for each cell across Cuba. Higher z values (yellow) correspond to greater inferred activity. As expected, the model highlights major urban centers like Havana and Santiago de Cuba, while rural interior regions show lower z, reconstructing a plausible economic landscape.



Estimated Economic Activity Across Cuba
(posterior mean of latent variable z)
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
Latent economic activity (z)

Figure 2: Estimated latent economic activity index z_i across Cuba.

Figure 3 presents the posterior standard deviation of the z_i, which is relatively low overall (0.04-0.06), indicating high confidence in estimates. Slightly higher uncertainty appears near Havana particularly, while remote areas exhibit uniformly low uncertainty.

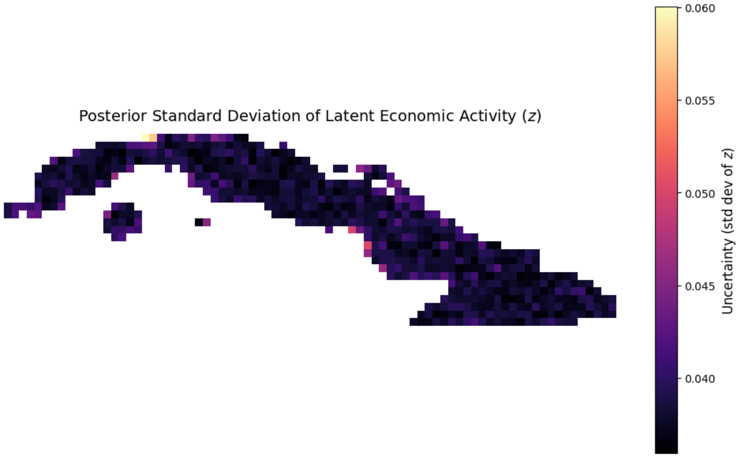

Posterior Standard Deviation of Latent Economic Activity (z)
0.060
0.055
0.050
0.045
0.040
Uncertainty (std dev of z)

Figure 3: Estimated latent economic activity index z_i across Cuba.

The learned coefficients reveal how each proxy contributes to z (Figure 4). The NDVI coefficeint is modest, with $\beta_{\text{NDVI}} \approx 0.143$, while roads have the strongest influence, with $\beta_{\text{roads}} \approx 2.289$. These values suggest that, in Cuba, road density is the most strongly correlated with the latent economic factor. NVDI has a small positive coefficient, implying that greener areas are slightly associated with higher economic activity when controlling for lights and roads, but the NVDI effect is quite weak: its 95% CI includes 0.

In terms of the noise-scale posteriors, the posterior mean for $\sigma_{\text{lights}} \approx 0.395$, $\sigma_{\text{NDVI}} \approx 1.082$, and $\sigma_{\text{roads}} \approx 0.080$. These suggest that nightlights are a fairly clean signal, while NVDI is much noiser and that greeness may contain a large amount of variation unrelated to economic activity. The road density appears noise-free, but it is important to note that the $\hat{R}$ is high (1.501), which could be explained by the extreme sparsity of the road raster.

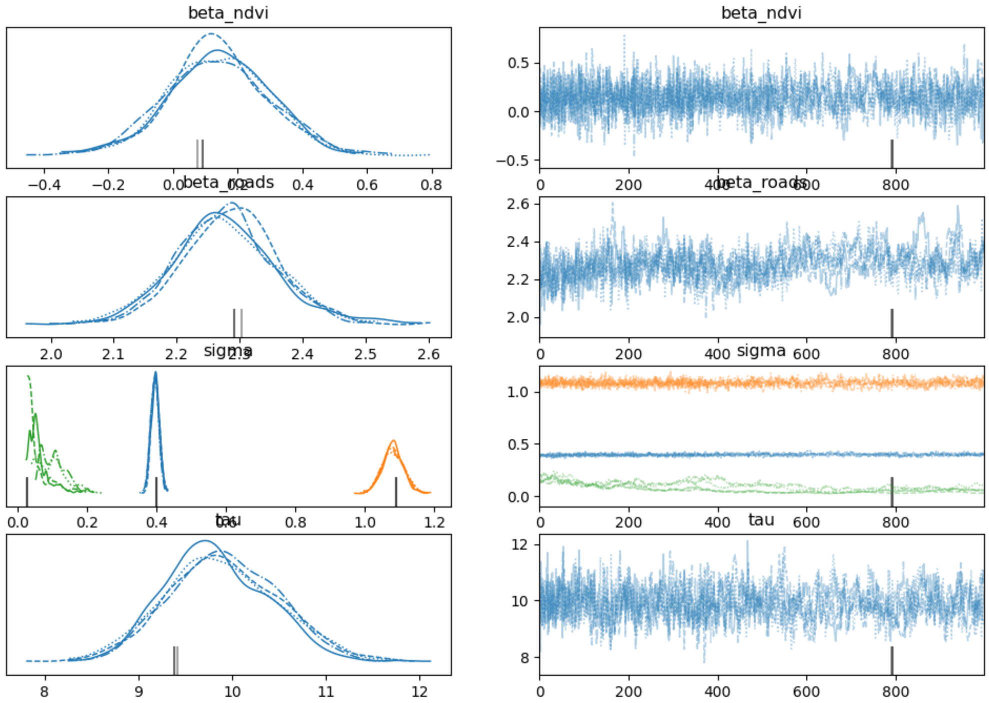


Figure 4: Marginal Posteriors (left) and MCMC Trace Plots (right) for the model parameters.

5 Multi-Resolution Analysis

I now refit the entire Bayesian model on three alternative spatial grids obtained by averaging the original 500 m rasters into square blocks of (i) 5 km, (ii) 10 km, and (iii) 20 km.

5.1 Proxy Coefficients across Resolutions

We observe two patterns emerge as we run the model across different resolutions.

First, β_{roads} declines from 4.440 at 5 km to 3.061 at 20 km, likely because coarse grids merge roadless and road-dense areas, diluting the relationship between roads and economic activity.

Second, and more interestingly, β_{NDVI} flips sign: negative at 5 km (-0.276), but positive at 20 km (0.0676). This may be an instance of Simpson's Paradox, where an aggregated trend reverses the direction observed in finer groups. At finer resolutions, NDVI may be negatively associated with development (e.g., forests in poor areas), but when averaged over large heterogeneous regions, greener areas might coincide with economically productive agricultural zones, flipping the overall correlation.

These shifts underscore the value of keeping coefficients global. Allowing β_k to vary by location would introduce many additional parameters, yet we only have three proxies per cell. We also have the ICAR prior to handle spatial smoothing, making global coefficients a reasonable and interpretable choice.

5.2 Spatial Smoothness

We also observe that the posterior τ drops from 17.389 to 7.023 with coarser grids, meaning that the model tolerates larger absolute differences between neighboring cells.

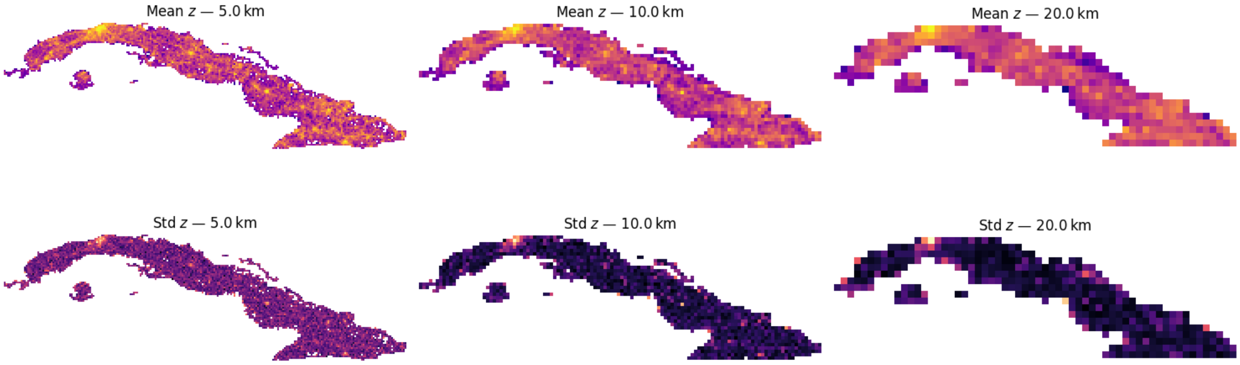


Figure 5: Posterior mean (top row) and posterior standard deviation (bottom row) of latent economic activity for 5 km, 10 km, and 20 km grids.

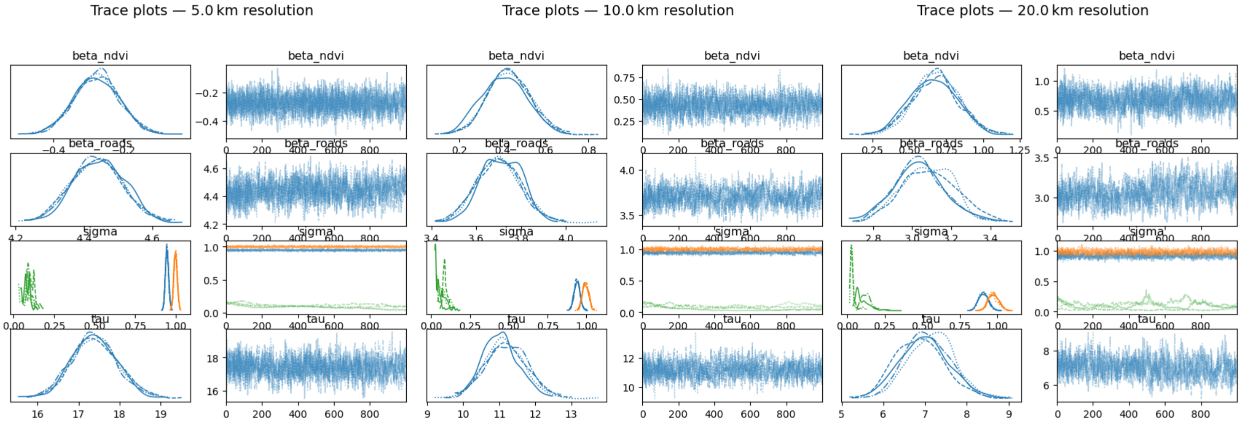


Figure 6: Trace plots at 5 km, 10 km, and 20 km.

6 Cross-Country Validation: Dominican Republic

To assess the external validity of my latent economic activity model, I now conduct a cross-country comparison using the Dominican Republic (DR) as a benchmark. The DR was chosen for three main reasons: (1) it shares geographic, environmental, and developmental similarities with Cuba, making it a reasonable comparison case; (2) it has better coverage of socioeconomic statistics and independent validation datasets; and (3) high-resolution data on wealth is publicly available, allowing spatially granular evaluation of the model outputs.

For validation, I use Meta's *Relative Wealth Index* (RWI), a dataset of wealth microestimates for 93 low- and middle-income countries (Cuba is not included!). The RWI is built from a combination of de-identified mobile phone connectivity data, satellite imagery, and machine learning techniques, and provides estimates of relative household wealth at a 2.4 km resolution across the DR (Chi et al. 2022). This data is independent of the three proxies used in my model, making the RWI an ideal source of ground-truth comparison.

First, I apply the same Bayesian latent variable framework to the DR using the three aligned raster proxies. After aggregating to the four different grid sizes and estimating the posterior mean latent index z, I map each RWI location's latitude and longitude to its corresponding coarse cell. Each RWI data point is then assigned the model-estimated z value at its location. After filtering to points with both a valid RWI and z, I calculate the correlation between the two series. The results are shown below in Figure 7.

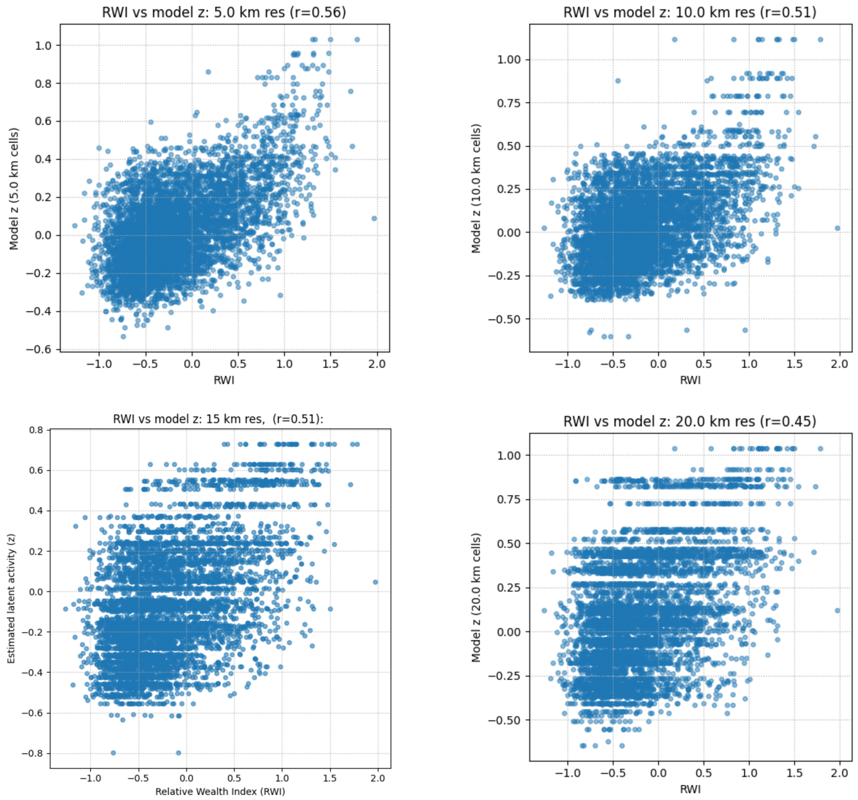


Figure 7: Comparison of model-estimated latent activity (z) with Meta's Relative Wealth Index (RWI) across spatial resolutions.

Overall, the validation results show a consistent positive correlation between the model's estimated latent economic activity and the RWI across all spatial resolutions. At a 5 km resolution, the correlation is strongest at $r = 0.56$. But as the resolution becomes coarser, the correlation gradually weakens, from $r = 0.51$ at both 10 km and 15 km, down to $r = 0.45$ at 20 km. This trend reflects the expected trade-off: coarser grids average over heterogeneous areas and smooth out localized wealth differences. Nevertheless, the moderate correlation even at 20 km affirms that the model captures meaningful economic signals.

Another interesting observation is the increasing hetereoskedasticity at finer spatial resolutions, particularly at 5 km and 10 km resolutions. As shown in Figure 7, the variance of predicted z values increases with higher RWI scores. This matches what we observe with our model's posterior standard deviation in wealthier areas like Greater Havana. The result reinforces the model's self-awareness about its limitations in high-wealth, heterogeneous environments.

7 Conclusion

In this project, I developed a Bayesian hierarchial model to infer a latent economic index of economic activity in Cuba from multiple proxy measurements. By combining nighttime light intensity, vegetation greener, and raod density within a single latent variable framework, I constructed a fine-grained economic activity map that is spatially coherent and plausible. The inclusion of an ICAR prior in the model enabled neighboring areas to have similar latent values, yielding smooth transitions that align with Cuba's known geographic and urban-rural patterns.

Importantly, we externally validated the approach. When applied to the DR, the inferred latent index showed a strong correspondence with Meta's RWI, a benchmark of local wealth. This validation in a data-rich context lends credibility to our Cuba results.

Despite these promising results, several limitations of our approach need to be acknowledged. First, the estimated latent economic index is on a relative scale; it is not anchored to absolute economic units like GDP or income, which means the values indicate relative differences but not the actual size of economic output. Second, our proxy data coverage is limited: nighttime lights, vegetation, and road density each capture only certain facets of economic activity, so sectors outside their scope (for example, informal economies) may not be well represented. Additionally, the use of a spatial smoothing prior introduces the risk of spatial confounding and potential over-smoothing, as previously discussed.

Future work includes: calibrating the latent index against ground-truth economic data where available, so that the index can be translated into absolute economic terms; incorporating additional or alternative proxies, such as mobile phone usage patterns, and detailed land use types, to broaden the evidence base for economic activity and improve coverage of hard-to-measure sectors; and finally, refining the model itself - perhaps allowing spatially varying proxy coefficients to account for regional differences.
Word Count: 1903

References

Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2022). Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences*, *119*(3), e2113658119. https://doi.org/10.1073/pnas.2113658119

Duncan, E. W., White, N. M., & Mengersen, K. (2017). Spatial smoothing in bayesian models: A comparison of weights matrix specifications and their impact on inference. *International Journal of Health Geographics*, *16*(1), 47. https://doi.org/10.1186/s12942-017-0120-x

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

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

Appendix: Github Repository

All code, data, and any scripts used to generate the results in this writeup are available at the following GitHub repository:

`https://github.com/cpez04/stat288-finalproject.`

To reproduce the results, please clone the repository and follow the instructions provided in the README.

Thank you for reading this far!