Detecting Data Misreporting From Space: Electricity In China

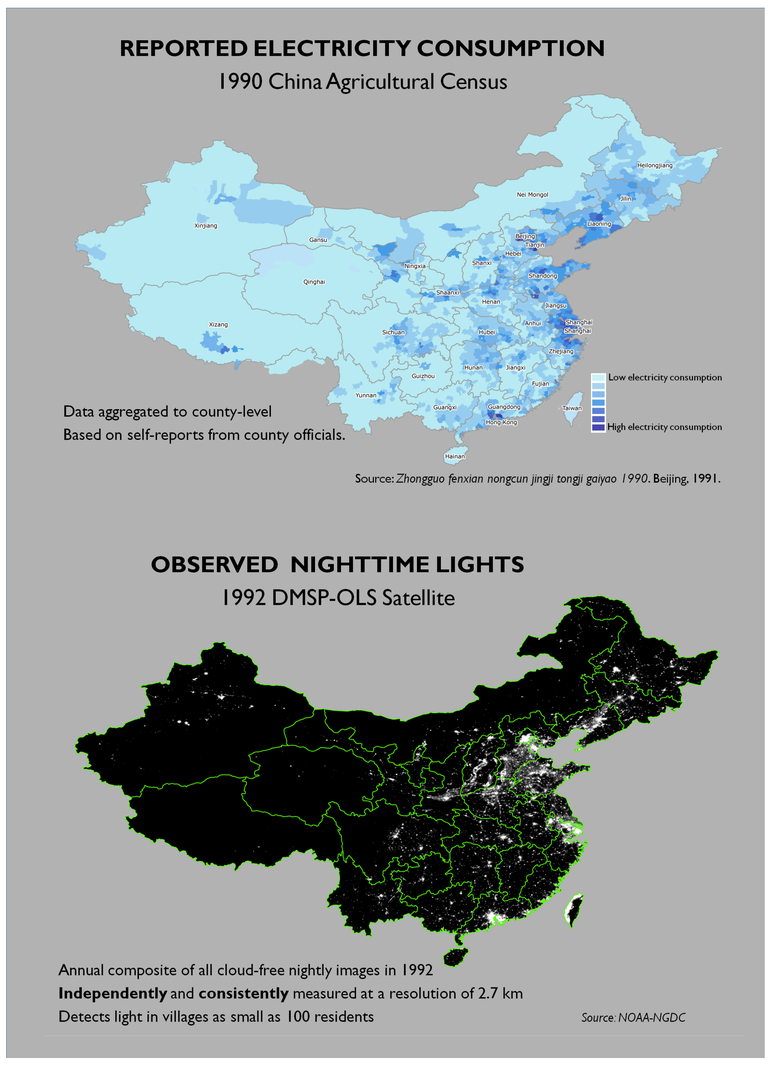
Patrick Lam[†] and **Brian Min**[‡]

[†]Department of Government, Harvard University (plam@fas.harvard.edu)
[‡]Department of Political Science, UCLA (bmin@ucla.edu)

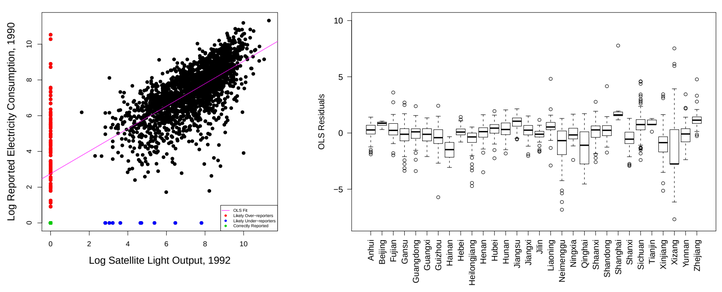
Summary

- The quality and accuracy of many measures of public service provision are unknown because they rely on government self reports with limited possibilities for external verification, especially in the developing world.
- We use satellite imagery of the earth at night to detect errors in officially reported electricity consumption data in 2300+ counties in China.
- We find evidence of substantial over-reporting of electricity, especially in Tibet (Xizang) and Sichuan.

Verifying Government Data on Electricity



A First Look



While there is a strong relationship between reported electricity consumption rates and nighttime light output, there are numerous outliers including **likely over-reporters** with zero satellite light output and **likely under-reporters** with zero reported electricity consumption. We assume that counties with both zero satellite light output and zero reported electricity consumption to be **correctly reported**.

The distribution of simple residuals from the OLS best fit line reveals geographic clustering of extreme over-reporters in provinces including Tibet (Xizang) and Sichuan.

Empirical Strategy

1. Estimate Actual Electricity Consumption.

1. Run a regression of electricity consumption, omitting **likely over-reporters** (zero satellite light output and non-zero electricity consumption) and **likely under-reporters** (non-zero satellite light output and zero electricity consumption):

$$\ln(\text{electricity}) = \ln(\text{satellite light output}) + \ln(\text{gross county output}) + \ln(\text{population}) + \ln(\text{area}) + \%\,\text{urbanization} + \%\,\text{industrial employment} + \%\,\text{agricultural employment} + \epsilon$$

2. Use predicted values from the regression as estimates of actual electricity consumption.

Problem: OLS is not robust to outliers, so data misreporting in electricity consumption (dependent variable) can bias our coefficients and estimates of actual electricity consumption.

Solution: Use robust least trimmed squares (LTS) estimator (Rousseeuw and Leroy, 1987). The LTS estimator minimizes the sum of the h smallest squared residuals (where $h > n/2$ and is defined by the user) and allows us to robustly estimate the relationship between electricity consumption and its predictors even in the presence of error or misreporting in up to 50% of the observations.

2. Identify Misreporting Counties.

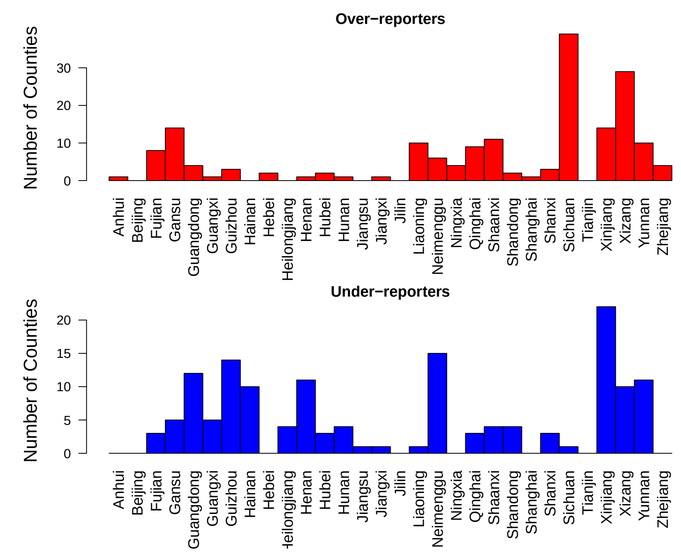
For observation i, the LTS standardized residual is

$$\frac{r_i}{\hat{\sigma}} = \frac{y_i - \hat{y}}{\hat{\sigma}}$$

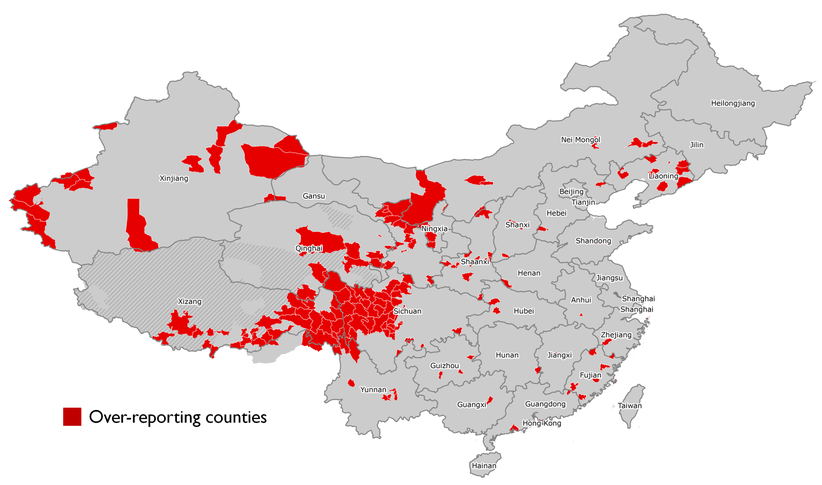
A county is an over-reporter if

- $\frac{r_i}{\hat{\sigma}} > 1.64$ (top 10% of distribution of outliers) OR
- the county is a **likely over-reporter** AND
- the county is not **correctly reported**

We follow a similar procedure to identify under-reporters.



Over-reporting counties are concentrated in western China

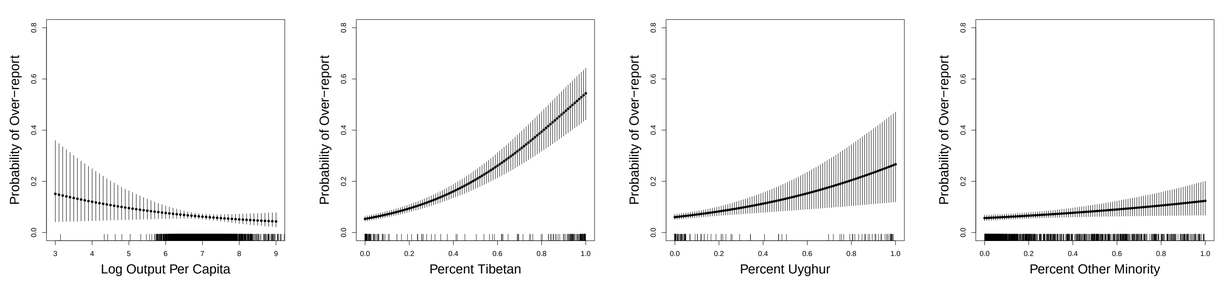


- Most over-reporters are concentrated in less developed Western China, especially in:
  - Tibet (Xizang)
  - Western Sichuan (home to a large Tibetan population)
  - Xinjiang (home to many Uyghurs)
- In politically sensitive minority regions, party officials may face even greater incenties to over-report performance.

3. Explain Over-Reporting.

- We expect government officials to be less constrained to report accurate data where the incentives for performance are high and where accountability to citizens is low.
- We test whether officials in poor counties and those with high minority populations are more likely to over-report using rare events logistic regression:

$$\text{Overreporting} = \ln(\text{output per capita}) + \text{percent Tibetan} + \text{percent Uyghur} + \text{percent other minority}$$



Results are probabilities of over-reporting and 95% confidence intervals, holding other variables at their means.