Modeling Spatial Imprecision with geoSIMEX

Robert Marty
Seth Goodman
Michael LeFew
Ariel BenYishay
Carrie Dolan
Dan Runfola

AidData

Institute for the Theory and Practice of International Relations

The College of William and Mary

danr@wm.edu
abenyishay@aiddata.org

geo.aiddata.org
geoquery.org

Summary

When information is geocoded from text documents - i.e., information on where an event occurred is assigned a latitude and longitude - spatial imprecision is a common challenge. For example, in the case of information about international aid extracted from donor reports, the district within which a well was built may be reported, but not the exact latitude and longitude of the well. The aim of this piece is to provide guidance to practitioners interested in using spatially-referenced information in cases where precise spatial boundaries or locations are not always known. We focus on this challenge in the context of development finance gathered by AidData using the *geo(query)* tool (geo.aiddata.org/query). We further provide software and a step-by-step example.

geoSIMEX

While the methods summarized in this document can be applied to many cases of spatial imprecision, we focus on an example application which leverages *geo(query)* (geo.aiddata.org/query), an online tool which enables quick access to integrated sources of spatial data for non-experts. We specifically provide guidance on how to apply the method detailed in this report, *geoSIMEX*, to data extracted from the *geo(query)* online tool for the purpose of causal inference. *geoSIMEX* is provided as an easy-to-use R package, and is designed explicitly for use with the *geo(query)* (http://geo.aiddata.org/query) tool. It leverages information on the imprecision in geographic measurements created by textual document geocoding to provide more accurate estimates of both impact effects (i.e., the impact a well had on agricultural productivity) as well as the certainty of these estimates.

Table of Contents

Summary
- geoSIMEX

Table of Contents

Summary of geoSIMEX

Quick Start
- Step 1: Download and Install R
- Step 2: Download and Install RStudio
- Step 3: Choose a Boundary using geo(query)
- Step 4: Choose Datasets using geo(query)
- Step 5: Check your Email and Download the Results
- Step 6: Launch R Studio and Install geoSIMEX and sp in R
- Step 7: Add your Data
- Step 8: Calculate the expected value of aid for each administrative division
- Step 9: Run a normal, linear model that does not account for Spatial Imprecision
- Step 10: Run a geoSIMEX model

Technical Details of geoSIMEX
- Introduction
- Definitions
- Probability Model with Example
  - Probability Aid Allocated to ROI
  - Expected Value of Aid in ROI
- Linear Regression with Spatial Uncertainty using geoSIMEX
  - Quantifying Spatial Uncertainty
  - geoSIMEX Steps
    - Step 1 - Estimate Naive Model
    - Step 2 - Simulating Additional Error
    - Step 3 - Binning
    - Step 4 - Coefficient Extrapolation
    - Step 5 - Standard Error Bootstrapping
    - Step 6 - Standard Error Calculation

Example of geoSIMEX with Simulated Data

Example of geoSIMEX using AidData

Acknowledgements

This work was made possible by the support of USAID, the Hewlett Foundation, KFW, Humanity United, the World Bank, the Global Environmental Facility, the MacArthur Foundation, and the College of William and Mary. We would also like to thank the many team members who contributed to this project, including Ben Dykstra and the AidData Data Team (Zhonghui Lv, Lauren Harrison, Alex Kappel, Sid Ghose, Brooke Russell). This work was performed in part using computational facilities at the College of William and Mary which were provided with the assistance of the National Science Foundation, the Virginia Port Authority, Virginia's Commonwealth Technology Research Fund and the Office of Naval Research.

About AidData

AidData is a research and innovation lab located at the College of William & Mary that seeks to make development finance more transparent, accountable, and effective. Users can track over $40 trillion in funding for development including remittances, foreign direct investment, aid, and most recently US private foundation flows all on a publicly accessible data portal on AidData.org. AidDta's work is made possible through funding from and partnerships with USAID, the World Bank, the Asian Development Bank, the African Development Bank, the Islamic Development Bank, the Open Aid Partnership, DFATD, the Hewlett Foundation, the Gates Foundation, Humanity United, and 20+ finance and planning ministries in Asia, Africa, and Latin America.

Recommended Citation

Marty, R., Goodman, S., LeFew, M., BenYishay, A., Runfola, D. 2016. *Modeling AidData Using geo(query).* AidData. Available online at http://geo.aiddata.org/docs/.

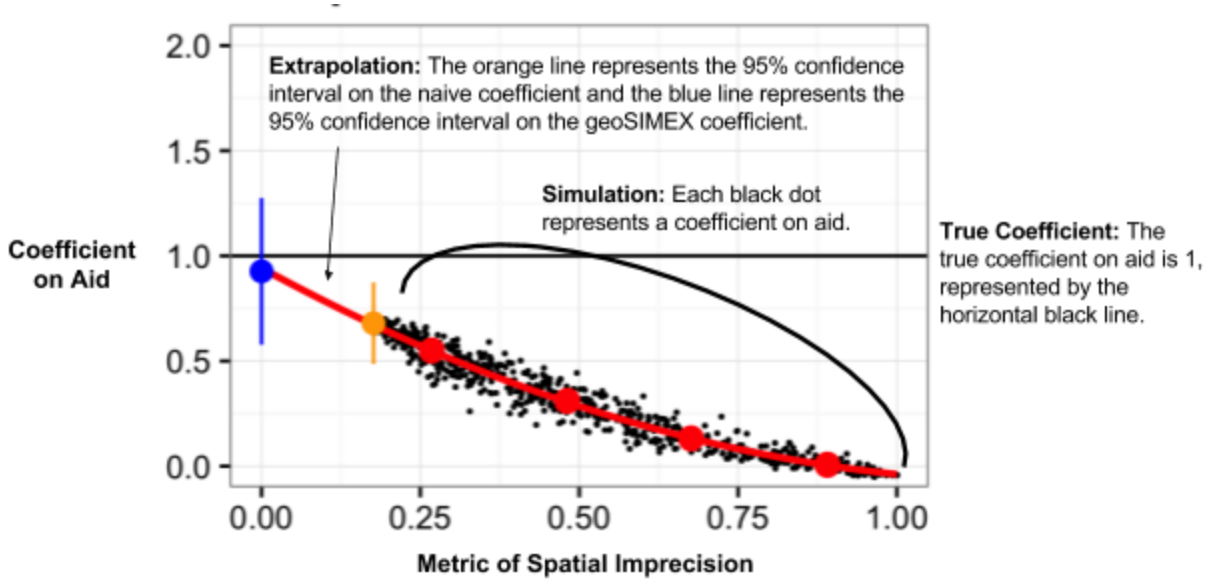
Summary of geoSIMEX

The Geographic Simulation and Extrapolation Method---geoSIMEX---is a method to correct for spatial imprecision in regression analyses. The process is based on two steps: simulation and extrapolation. The simulation step involves estimating regressions with simulated additional imprecision. Here, a relationship in established between regression coefficients and the magnitude of spatial imprecision. In the extrapolation step, the trendline between regression coefficients and spatial imprecision is used to extrapolate the coefficients back to a point of no spatial imprecision.

Simulation: In the simulation step, additional spatial uncertainty is simulated in the data. For example, if project documentation indicates that aid was allocated to one of ten subcounties in a district, simulating could involve specifying that aid was allocated to one of thirty subcounties within a region or one of one-hundred subcounties within a country. In the simulation step, the spatial imprecision of each aid project is increased by a random amount. A regression is then estimated using the simulated data. The aid variable is constructed by taking an expected value of aid in each district. For example, if a $10 million aid project was allocated to one of 10 subcounties, each subcounty would be assigned $1 million from that aid project (users can also specify a probability of a subcounty receiving aid, and aid will be allocated proportionally to that probability). Within each simulation, the primary regression of interest is then estimated and coefficient values collected. This process is repeated 500 times with varying degrees of spatial imprecision introduced in each simulation.

Extrapolation: The simulation step generates regression coefficients from 500 models and a metric of spatial imprecision associated with each model. Using this data, a trend is fit between the regression coefficients and the level of uncertainty. Using this trend, a value of the regression coefficient is extrapolated back to a point of zero spatial imprecision.

The below figures illustrates the geoSIMEX process. In the below example, the relation between aid and some outcome metric (e.g., wealth) is estimated. The true coefficient on aid is one. The researcher receives spatially imprecise data and estimates a naive model ignoring spatial imprecision (the orange line represents the 95% confidence interval of the naive model). The naive model does not capture the true coefficient on aid. The black dots to the right of the naive confidence interval represent aid coefficients on models using data with simulated additional spatial uncertainty. The red line is the trend line between the coefficient and spatial uncertainty, and is extrapolated back to a point of zero spatial imprecision. The geoSIMEX aid coefficient (represented in blue) is formed at this point of zero spatial imprecision.



geoSIMEX: Simulation and Extrapolation

Quick Start

This quick start guide will introduce you to a method for using data on international aid from AidData in modeling efforts. Frequently, when text sources are used to identify the geographic location of where something happened (e.g., where a well was built to supply water for agricultural purposes), the exact latitude and longitude at which the event occurred is not known with complete precision (see Figure 1). This is challenging to researchers who seek to analyze the impact of international aid on relevant outcome measures (e.g., agricultural productivity), as this imprecision - if not incorporated into the modeling process - can lead to bias in estimated causal effects. This section provides a brief quick-start guide for users that are interested in learning how to leverage the *geo(query)* online toolkit for merging information on international aid with relevant satellite (and other) outcome and covariate data, and then model this data using the *geoSIMEX* tool in a way that accounts for spatial imprecision.

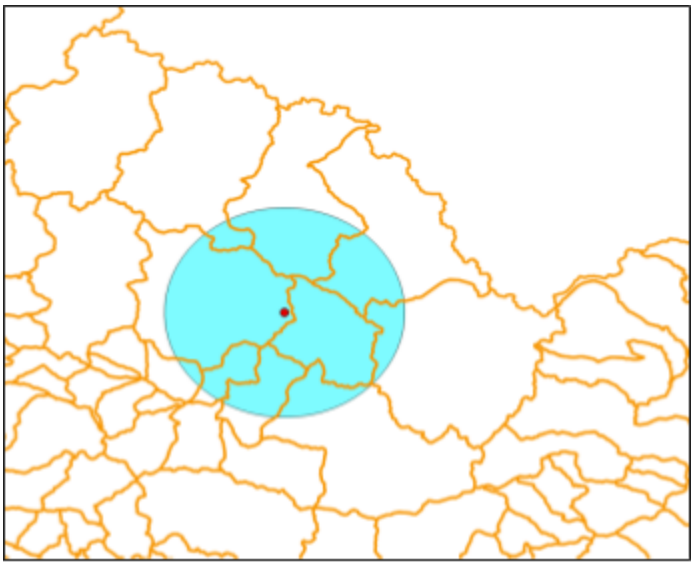


Figure 1. Geographic Area to which aid is known to have gone (blue).

Figure 2. The geo.query boundary interface.

Step 1: Download and Install R

While this document is written for individuals with no expertise in coding, the *geoSIMEX* tool is written in the R programming language, an open source, freely available language. As a first step, you will need to download and install R using the links for your operating system found at https://cran.rstudio.com/ .

Step 2: Download and Install RStudio

As a second step you should download RStudio which provides a user interface that will be more familiar to users that have used other statistical programs (e.g., SPSS, STATA). While you can choose variable levels of support, the free, open source version required for this guide can be downloaded from https://www.rstudio.com/products/rstudio/download/

Step 3: Choose a Boundary using *geo(query)*

Go to http://geo.aiddata.org/query , which is the website that provides both (a) international aid data aggregated to variable geographic scales, and (b) a wide variety of relevant satellite, social, environmental, and other covariate data. The first page will prompt you to choose a boundary, as seen in Figure 2. Here, "GADM" is a description of the source of the geographic boundaries (http://www.gadm.org). Type "Uganda" and choose ADM3 boundary, which will update the map to show all counties in Uganda (which will be our unit of analysis for this example). Click "Search Datasets" to move on to the next page.

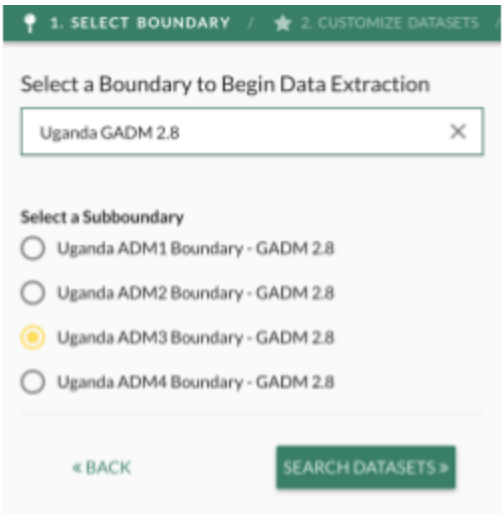


Figure 3. Administrative Boundary Selection.

Step 4: Choose Datasets using *geo(query)*

After clicking "Search Datasets," select the following datasets (these will appear on the "datasets" column to the left).

- **Uganda Geocoded Aid Data v1.4.1:** Under sector names, select "Government and civil society, general." Click "create more filters" then "years," and select all years from 2000-2010. Then, click "+ADD TO REQUEST."
- **Nighttime Lights:** Select "mean" and 2010. Then, click "+ADD TO REQUEST."
- **Population (GPW V3, UN Adjusted):** Select "Sum" and 2000. Then, click "+ADD TO REQUEST."
- **Conflict Events:** Select "Sum." Then, click "+ADD TO REQUEST."
- To finish, select "Submit Request" at the top right-hand-side of the page, then click "SUBMIT REQUEST" again. Add your email, then click "SUBMIT."

Step 5: Check your Email and Download the Results

Soon you'll receive an email saying that your request has processed via email. Download and open the zip-file from the email. To help follow along in the remaining steps, change the name of:

- the folder to "geoSIMEX_example"
- the .csv file in the folder to "uganda_data.csv,"

Step 6: Launch R Studio and Install geoSIMEX and sp in R

You'll need a few R packages, which you install from inside of R: (1) sp allows R to read spatial data, (2) raster provides administrative level data, (3) jsonlite allows R to read json files, and (4) geoSIMEX allows modeling the information you downloaded from geo(query) in a way that incorporates spatial imprecision. Devtools is used to load R packages from github, where the most up-to-date version of geoSIMEX - along with all past versions - is maintained.

To load the packages, type in the text as shown below then click "Run."

```
install.packages("devtools")
install.packages("sp")
install.packages("jsonlite")
install.packages("raster")
install.packages("rvest")
install_github("itpir/geoSIMEX")
library(geoSIMEX)
library(sp)
library(jsonlite)
library(raster)
library(rvest)
```

Step 7: Add your Data

You'll need to import two files from the geo(query) download in order to use geoSIMEX.

First, you'll need country administrative level data which comes from a CSV. You can import the CSV using the "read.csv" function, using the full file path to the csv file:

```
uganda_data <- read.csv("~/Desktop/geoSIMEX_example/uganda_data.csv")
```

Second, we'll rename some of the variables in the administrative level data:

```
names(uganda_data)[names(uganda_data) == "gpw_v3_count.2000.sum"] <- "population"
names(uganda_data)[names(uganda_data) == "v4composites_calibrated.2000.mean"] <- "NTL"
names(uganda_data)[names(uganda_data) == "acled_v3_count.none.sum"] <- "conflict"
```

Third, one of the population values is missing. We set this missing value to zero:

```
uganda_data$population[is.na(uganda_data$population)] <- 0
```

Fourth, you'll need aid level information. The geo(query) download contains a JSON[1] file, which includes information on how you filtered the aid information. Here, we'll import the JSON file and use it to subset the aid information based on the filter information.

```
uganda_json <- fromJSON("~/Desktop/geoSIMEX_example/request_details.json")
```

Next, we'll use the JSON file and the "subset.aiddata" function to create an aid dataset that is subsetted according to the specified filters in geo(query). The subset.aiddata function downloads the aid-level information from online and filters it according to the parameters in the JSON file.

```
uganda_aid <- subset.aiddata(uganda_json)
```

Step 8: Calculate the expected value of aid for each administrative division

The expected_aid_ROI function calculates the expected value of aid. Below are the parameters for the function:

- **aidData** is the name of the aid level dataframe
- **aid.project.amount** is the name of the variable in the aid level dataset that contains aid amount information
- **aid.precision.code** is the name of the variable in the aid level dataframe indicating the precision code of aid projects.
- **aid.pc1.centroid.name** is the name of the variable in the aid dataset that corresponds with the administrative zone that the aid project's coordinates (latitude and longitude) fall in. Here, the administrative zone should be the same level as the precision code 1 administrative level (i.e., here, we associate precision code 1s with ADM3 (NAME_3); consequently, we use ADM3 (NAME_3) for aid.pc1.centroid.name.
- **roiData** is the name of the administrative division level dataframe
- **roi.prob.aid** is the name of the variable in the administrative division dataframe indicating how we will allocate spatially imprecise aid. Here, we use population.
- **roi.pc#.name** is the name of the variable in the administrative level dataframe that corresponds with the precision code. For example, roi.pc3.name="NAME_1" means we are associating precision code three with the administrative zone capture in the variable NAME_1, which is ADM1.

```
uganda_data$expected_aid <- expected_aid_ROI(aidData=uganda_aid,
                                             aid.project.amount="total_commitments",
                                             aid.precision.code="precision_code",
                                             aid.pc1.centroid.name="NAME_3",
                                             roiData=uganda_data,
                                             roi.prob.aid="population",
                                             roi.pc1.name="NAME_3",
                                             roi.pc2.name="NAME_2",
                                             roi.pc3.name="NAME_1",
                                             roi.pc4.name="NAME_1",
                                             roi.pc5.name="NAME_1",
                                             roi.pc6.name="NAME_0")
```

Step 9: Run a normal, linear model that does not account for Spatial Imprecision

To run a simple OLS model, type the following into R. Here, we seek to uncover the relation between aid and conflict. The number of conflict events in each administrative division is our dependent variable, and aid is out independent variable of interest. We also control for nighttime lights—which is used as a proxy for local economic development— in 2000.

Run an OLS model using conflict as the dependent variable and aid and nighttime lights as independent variables.[2]

[1] A JSON (JavaScript Object Notation) file that is readable by many programming languages that is often used to store data or information about datasets.

[2] When interested in causally identifying the impact of aid, the matchIt algorithm can be run. The function will subset the administrative level division data so that specified covariates are statistically similar across treatment and control groups. For more information on geoMatch, see: http://geo.aiddata.org/docs/geoMatch.pdf

```
naive_model <- lm(conflict ~ expected_aid + NTL, data=uganda_data)
```

To view regression results of the naive model, type:

```
summary(naive_model)
```

The summary function will display the following results, which includes parameter estimates. The below results indicate that aid is having a statistically significant impact on increasing the number of conflict events. Step 10 provides more detail on the interpretation of the values returned by the summary function.

```
> summary(naive_model)

Call:
lm(formula = conflict ~ expected_aid + NTL, data = uganda_data)

Residuals:
    Min      1Q  Median      3Q     Max
-21.756  -3.241  -2.389  -0.962 162.574

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.344e+00  4.624e-01   5.070 4.78e-07 ***
expected_aid 1.280e-07  2.602e-08   4.918 1.03e-06 ***
NTL          7.159e-01  8.118e-02   8.819  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.7 on 963 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.09969,   Adjusted R-squared:  0.09782
F-statistic: 53.31 on 2 and 963 DF,  p-value: < 2.2e-16
```

Step 10: Run a geoSIMEX model

The following are parameters in the geoSIMEX function:

- **model** is the naive model that does not account for spatial imprecision (estimated in step 8)
- **geoSIMEXvariable** is the name of the variable in the naive model that is measured with spatial uncertainty
- **aidData** is the name of the aid level dataframe
- **aid.project.amount** is the name of the variable in the aid level dataset that contains aid amount information
- **aid.precision.code** is the name of the variable in the aid level dataframe indicating the precision code of aid projects.
- **aid.pc1.centroid.name** is the name of the variable in the aid dataset that corresponds with the administrative zone that the aid project’s coordinates (latitude and longitude) fall in. Here, the administrative zone should be the same level as the precision code 1 administrative level (i.e., here, we associate precision code 1s with ADM3 (NAME_3); consequently, we use ADM3 (NAME_3) for aid.pc1.centroid.name.
- **roiData** is the name of the administrative division level dataframe
- **roi.prob.aid** is the name of the variable in the administrative division dataframe indicating how we will allocate spatially imprecise aid. Here, we use the spatial area.
- **roi.pc#.name** is the name of the variable in the administrative level dataframe that corresponds with the precision code. For example, roi.pc3.name=”NAME_1” means we are associating precision code three with the administrative zone capture in the variable NAME_1, which is ADM1.

```
geoSIMEX_model <- geoSIMEX(model = naive_model,
                           geoSIMEXvariable = "expected_aid",
                           aidData = uganda_aid,
```

```
aid.project.amount = "total_commitments",
aid.precision.code = "precision_code",
aid.pc1.centroid.name="NAME_3",
roiData = uganda_data,
roi.prob.aid="population",
roi.pc1.name="NAME_3",
roi.pc2.name="NAME_2",
roi.pc3.name="NAME_1",
roi.pc4.name="NAME_1",
roi.pc5.name="NAME_1",
roi.pc6.name="NAME_0")
```

To view regression results, type:

```
summary(geoSIMEX_model)
```

The summary function will display the following result, which includes coefficient estimates, standard errors, t-values, p-values, and significance levels. The below text explains the output.[3] The exact values for each user will not exactly match what is shown below; this is because geoSIMEX involves a random simulation step, where differences in results between users occur due to randomness. However, the magnitude of the estimates should be similar.

Here, no statistically significant relation is found between aid and the number of conflict events.

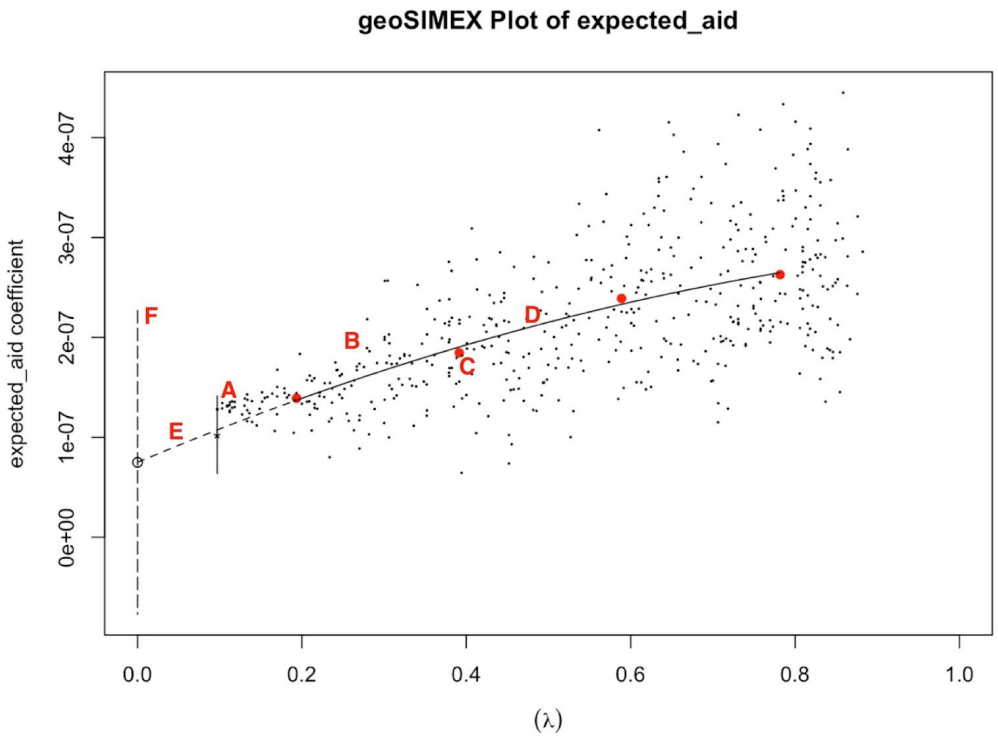
```
> summary(geoSIMEX_model)
Call:A
lm(formula = conflict ~ expected_aid + NTL, data = uganda_data)

             B   Estimate C Std. Error D t.value E  p.value F
(Intercept)  2.8433e+00 9.5382e-01  2.9810  0.002945 **
expected_aid 7.4956e-08 9.2345e-08  0.8117  0.417167
NTL          7.4301e-01 1.0315e-01  7.2032 1.185e-12 ***
---
Signif. codes: G 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **A. Call -** This is the regression equation, indicating the dependent variable, independent variables, and the dataset.
- **B.** This column shows the independent variables in the regression.
- **C. Estimate -** These are the coefficient estimates in the regression equation.
- **D. Std. Error -** These are the standard errors associated with the coefficients.
- **E. t.value -** These are the t-statistics associated with the coefficients, and are used to test whether the coefficients are statistically different from zero.
- **F. p.value -** These are the p-values associated with the coefficients, and show the probability of obtaining the coefficient (or one larger in magnitude) if the true coefficient was zero.
- **G. Signif. codes -** These denote levels of significance (i.e., 1% level, 5% level, etc), and are shown to the right of the p-values.

The plot function produces a figure with a metric of spatial uncertainty on the x-axis (lambda) and the variable coefficient on the y-axis. The plot function requires two inputs: the geoSIMEX model name and the name of the variable to plot. The below text explains the elements in the plot.

```
plot(geoSIMEX_model, variable="expected_aid")
```

[3] Language adapted from: http://www.ats.ucla.edu/stat/stata/output/reg_output.htm



- **A.** This is the 95% confidence interval of the coefficient in the naive model (i.e., the model that ignores spatial certainty).
- **B.** geoSIMEX estimates hundreds of models with simulated additional spatial uncertainty. The black points are values of the aid coefficient from these models.
- **C.** The red points are average values of coefficients within certain ranges of lambda, where lambda is a metric of spatial imprecision (a value of zero indicates no spatial imprecision, while a value of 1 indicates significant spatial imprecision).
- **D.** This is a quadratic trend line build on the average coefficient values (points from B).
- **E.** This is an extrapolated portion of the trend line from C.
- **F.** This is the 95% confidence interval of the coefficient on aid from the geoSIMEX model.

Technical Details of geoSIMEX

Introduction

This section presents a methodology for incorporating spatial imprecision into models which seek to estimate the impact of international aid (or other forms of development finance), geoSIMEX. This approach leverages known spatial imprecision in datasets to back-extrapolate to model solutions you would find under conditions of no spatial imprecision, and re-calculates uncertainty (i.e., standard errors) to account for this process. As our simulated and example cases illustrate, this results in more accurate representations of what claims can be made when accounting for the imprecision of the source data.

In studies of the effect of international development projects, the exact location of project implementation is often unknown, creating a source of uncertainty and concomitant difficulty in accurately ascribing effects to projects (see figure 1). This spatial uncertainty frequently stems from insufficient project documentation. At AidData, when international aid projects are geocoded, locations are assigned a precision-code which describes this spatial imprecision. We describe a method to calculate the probability that an aid project was located in a region of interest (e.g. districts in a country), given the known uncertainties in the data. We use this probability to (a) calculate the expected value of aid flowing to any area of interest, and (b) demonstrate how a probability model can be used to inform linear regression models through a geographic simulation and extrapolation (geoSIMEX) procedure.

Definitions

Location: A single location of a development finance project, with the possibility of multiple locations associated with a project. For example, a development finance project may entail the construction of multiple schools.

Region of Interest (ROI): An instance of our chosen unit of observation. For example, if our unit of observation is a district, then each district in a country is a unique region of interest.

Coverage: The area in which a project location may exist. For example, a project location may have a precision code and geographic information that tells us that a school was built somewhere within a district; in this case, the district would be the coverage area for that project location.

Overlap: The overlap of the area of coverage with the region of interest.

Probability Model with Example

The goal of this section is to describe the method used to calculate the probability that a particular project location falls in a given ROI (where a Region of Interest could be, for example, a district in Nepal), and the expected value of aid within an ROI.

Probability Aid Allocated to ROI

Each project location is assigned a set of coordinates and a precision code indicating the exactness of our knowledge of the location. Based on the precision code, we construct an area of coverage over which the project could be located. For instance, if a project location has precision code 1, the area of coverage is very small, i.e. within a small administrative zone or buffer. If a project location has a precision code 6 or 8, the area of coverage is an entire country. If there is some amount of overlap between the area of coverage of a location and a Region of Interest (ROI), we consider there to be some positive probability that the project location is in the ROI (see Figure 1). We assume that the probability that a project was allocated to an ROI is equal to the size of the area of overlap divided by the size of the area of coverage, as this is a conservative application of the available information (i.e., we make no assumptions regarding the spatial allocation of a project). The method presented here is extensible to cases in which the researcher chooses to make additional assumptions about how projects are spatially allocated (i.e., assuming aid is more likely to be allocated in areas with high populations), but we do not detail such an approach in this piece.

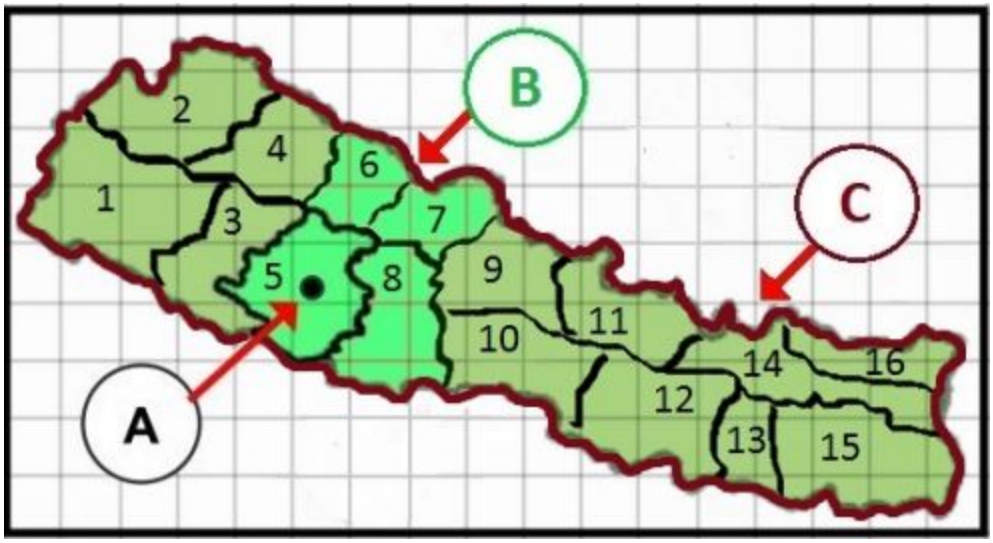


Figure 2. Project Areas

In figure 2, we present a hypothetical country with sixteen districts, which we will define as our sixteen regions of interest (ROIs). Districts 5,6,7, and 8 are distinguished on the map to illustrate the example. Blocks of the grid are each 2500 km2 . A hypothetical data set contains three geocoded international aid project locations of various levels of precision, locations A, B, and C. Location A is assigned a coordinate pair in District 5 and had strong documentation, resulting in precise geographic information (i.e. an exact latitude and longitude, or precision code 1). Due to weaker project documentation, location B has precision level indicating that it was allocated anywhere in the region that includes districts 5,6,7, and 8, denoted precision code 4. Project C has very uncertain spatial information, such that it may be anywhere in the country, denoted precision code 8. The area in square kilometers of the region in which each project location may have been allocated gives us the size of the project location's area of coverage (see table 1).

Using the area of coverage along with the area of each ROI, we calculate the probability that each location is in each ROI. Tables 1-3 demonstrate the calculation, where the probability that each project is in a district is the area of overlap divided by the area of coverage. Project A has area of coverage entirely within District 5, and therefore only overlaps with District 5. In this case, we have exact geographic information, reflected by the 100% probability of being in District 5. Project B has area of coverage over Districts 5-8, so the area of overlap for each of these districts is equal to the district's area, and zero for all other districts. Hence, the probability that Project B exists in each of districts 5-8 is proportional to the area of the district, and zero for all other districts. Because project C has area of coverage over the entire country, the area of overlap for all districts is equal to the district's area, and all districts have a positive probability of containing project C.

Project Location	Precision	Area of Coverage
A	1	Populated Area in District 5 (3 km2)
B	4	Districts 5-8 (32,000 km2)
C	8	Entire Country (112,500 km2)

Table 1. Example geocoded aid locations and associated geographic precision.

District	5	6	7	8	All Other
Area (km2)	*10,000*	*5,000*	*7,500*	*10,000*	*80,000*

Table 2. Geographic Area of Example Districts.

Project Location	Area of Coverage (km^2)	Overlap with District (km^2)					Probability in District				
		5	6	7	8	All Other	5	6	7	8	All Other
A	3	3	0	0	0	0	1	0	0	0	0
B	32,000	10,000	5,000	7,500	10,000	0	4/13	2/13	3/13	4/13	0
C	112,500	10,000	5,000	7,500	10,000	80,000	4/45	2/45	3/45	4/45	32/45

Table 3. Estimated probability that a given aid project falls into a region, given no additional information other than geographic size.

Expected Value of Aid in ROI

Using these probabilities we can calculate the expected value of aid within each ROI. The expected value is calculated by multiplying the dollar value of an aid project by the probability of the project falling into each ROI. This process is repeated for each aid project, and values are summed across ROIs to create the expected value of aid within each ROI. Table 4 illustrates calculating the expected value of aid for districts 5 through 8 from the above example. Here, we assume that $5 million was allocated through project A, $10 million through project B and $30 million through project C.

	Project A			Project B			Project C			Expected Value of Aid
	Prob.	$	**E[Aid]**	Prob.	$	**E[Aid]**	Prob.	$	**E[Aid]**	Within Districts
District 5	1	5	**5**	4/13	10	**3.07**	4/45	30	**2.66**	$5 + 3.07 + 2.66 = 10.73$
District 6	0	5	**0**	2/13	10	**1.53**	2/45	30	**1.33**	$0 + 1.53 + 1.33 = 2.86$
District 7	0	5	**0**	3/13	10	**2.30**	3/45	30	**2**	$0 + 2.30 + 2 = 4.30$
District 8	0	5	**0**	4/13	10	**3.07**	4/45	30	**2.66**	$0 + 3.07 + 2.66 = 5.73$

Table 4. Estimated value of aid allocated to a set of hypothetical regions of interest. Prob. refers to probability of the project being located in an ROI, $ refers to the total project amount in millions, and E[Aid] refers to expected aid in a given ROI (calculated by Prob. × $).

Linear Regression with Spatial Uncertainty using geoSIMEX

geoSIMEX builds on a general approach to measurement error, SIMEX. SIMEX leverages the relationship between increasing measurement error and bias following a two step process. First, SIMEX simulates additional measurement error to establish a relation between measurement error and covariate bias. Second, it uses the relation between error and bias to extrapolate to a point with zero measurement error, thus providing an estimate of the unbiased coefficient.

Quantifying Spatial Uncertainty

In other domains, SIMEX simulates additional measurement error by scaling the distribution of measurement error by some value. However, simulating additional measurement error through scaling a single, defiend distribution does not adequately capture increasing spatial uncertainty across potentially arbitrary sets of spatially-configured units. In our context, greater uncertainty is reflected by increasing the area of coverage of a project, thus expanding the set of ROIs where the project could be located. To reflect uncertainty driven by the extent of projects' area of coverage, we quantify spatial uncertainty (g) using:

$$\lambda = \frac{\sum_i^P Area\ of\ Coverage_i}{\sum_i^P Total\ Possible\ Area\ of\ Coverage_i} \qquad \text{eq. 1}$$

where i is an individual project out of P total projects. Area of Coverage is the known area of coverage for a given project *i* defined by its associated precision code - i.e., in the case of AidData the spatial area across which a project could be located. Total Possible Area of Coverage is the area of coverage of project i under complete spatial uncertainty - i.e., the case where no spatial information is known, or precision code 8 following the AidData schema.

If the latitude and longitude of every aid project was known, g would resolve to 0—indicating zero spatial uncertainty. If spatial data resolution was only available for the entire study area (e.g., aid provided for general budget support without indication of where the project was allocated), g would resolve to 1—indicating 100% spatial uncertainty.

geoSIMEX Steps

geoSIMEX is conducted in six discrete steps. Figure 3 illustrates the steps, and the subsequent paragraphs describe the geoSIMEX process. The figure and descriptions illustrate an example estimating:

$$\textbf{\textit{wealth}} = \Theta * \textbf{\textit{aid}} + \varepsilon \qquad \text{eq. 2}$$

where aid is causally and positively related to some wealth measure (e.g., aid in the form of unconditional cash transfers and wealth representing average dollar amount spent among individuals in an ROI), and ε is a random error term. Aid is measured with spatial uncertainty, and through using geoSIMEX we account for the spatial uncertainty to accurately estimate the model coefficient, θ.

Step 1 - Estimate Naive Model

The naive model is estimated where spatial uncertainty in the aid variable is ignored. The value of aid is the expected value of aid within each ROI, calculated following the steps described earlier. In figure 3a, the orange line represents the 95% confidence interval of the coefficient on aid in the naive model. The level of spatial uncertainty (g) of this model is about 0.4. The horizontal black line represents the true model coefficient, which the naive model fails to capture.

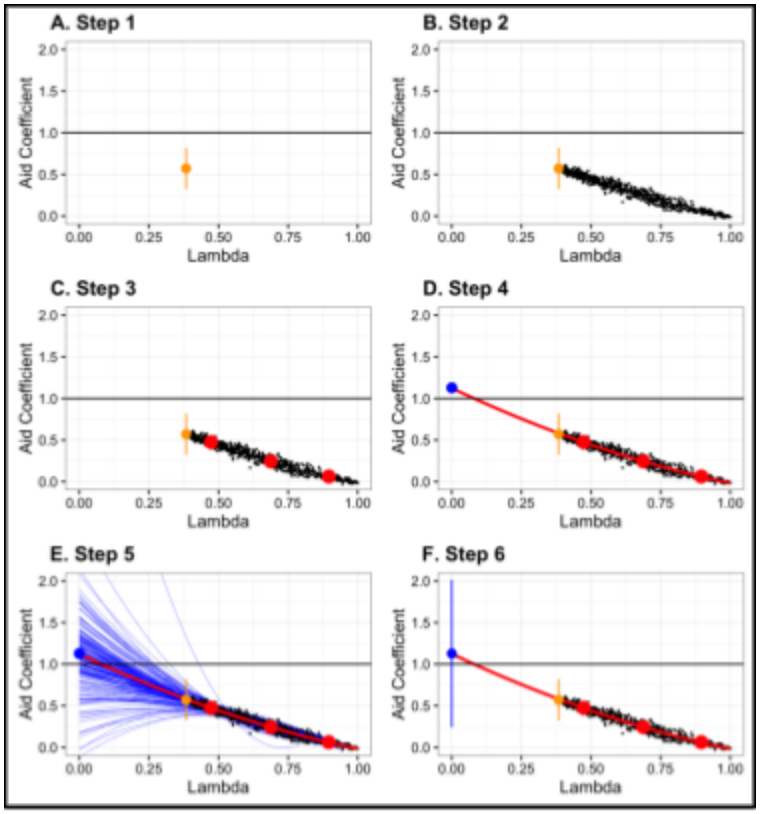


Figure 3. Example steps of the geo.simex process.

Step 2 - Simulating Additional Error

Additional uncertainty is simulated by randomly increasing degrees of spatial uncertainty to projects. For example, a project that has information with an exact latitude and longitude will randomly be assigned to a county, state, or even the entire country. At each iteration, the expected value of aid is calculated for each ROI. A model is fit using the aid variable with additional spatial uncertainty. This process is repeated 1000 times, where g (measure of spatial uncertainty) and the model coefficient, ɑ , are collected at each iteration. In figure 3b, the black points represent the model coefficients and their associated g values.

Step 3 - Binning

Model coefficients are separated into three equally-sized bins based on the level of spatial uncertainty (g) of the aid variable (e.g., if g values range from 0.4 to 1, coefficients are separated into bins of 0.4-0.6, 0.6-0.8, and 0.8-1). Average coefficient and g values are calculated within each bin, represented as red dots in figure 3c.

Step 4 - Coefficient Extrapolation

A quadratic trend is fit on the resulting average coefficient and lambda values. The trend is then extrapolated back to g = 0, thus providing an estimate of ɑ under complete spatial certainty. In figure 3d, the red line represents the extrapolated trend, and the blue dot represents the extrapolated estimate of the coefficient on aid.

Step 5 - Standard Error Bootstrapping

We employ a bootstrapping method to calculate standard errors of coefficients. Here, a point from each bin is sampled, a quadratic trend is fit on the resulting values, and the trend is extrapolated back to g = 0. This process is repeated 1000 times. In figure 3e, each blue line represents one extrapolated trend and estimate.

Step 6 - Standard Error Calculation

The standard deviation of the resulting values from the bootstrapping procedure becomes the standard error of the mean estimate. In figure 3f, the blue line represents the 95% confidence interval. Here, the geoSIMEX estimate captures the true coefficient.

Example of geoSIMEX with Simulated Data

This section provides an example using geoSIMEX. We estimate the relation between aid and wealth described in equation 2, where aid is measured with spatial uncertainty. Figure 4 illustrates the estimated coefficient of aid on the Y-axis (ɑ), amount of spatial imprecision (g) on the X-axis, and the true coefficient (1) by a black horizontal line. In the first example (figure 4a),

there is relatively less certainty in the allocation of aid in the source dataset, represented by the minimum g of 0.4. In the second example (figure 4b), there is relatively more certainty in the source data, represented by minimum $g = 0.25$. In both cases, the linear model estimating aid is fit using the initial data ($g = 0.4$ and 0.25, respectively), and the coefficient estimate and 95% confidence intervals are visualized using the orange line. Each black point on the graph represents a geoSIMEX iteration, the red points represent average values within the three bins, and the red line represents the extrapolated estimate to the coefficient on aid when $g = 0$. Finally, the blue lines represent a bootstrapped estimate of the standard error based on the geoSIMEX procedure. These figures provide example illustrations of cases in which SIMEX can effectively mitigate bias in traditional linear models and provide more accurate representations of uncertainty as the quality of data decreases (represented by the larger confidence interval in the case of $g = 0.4$, contrasted to the smaller confidence interval in the case of $g = 0.25$).

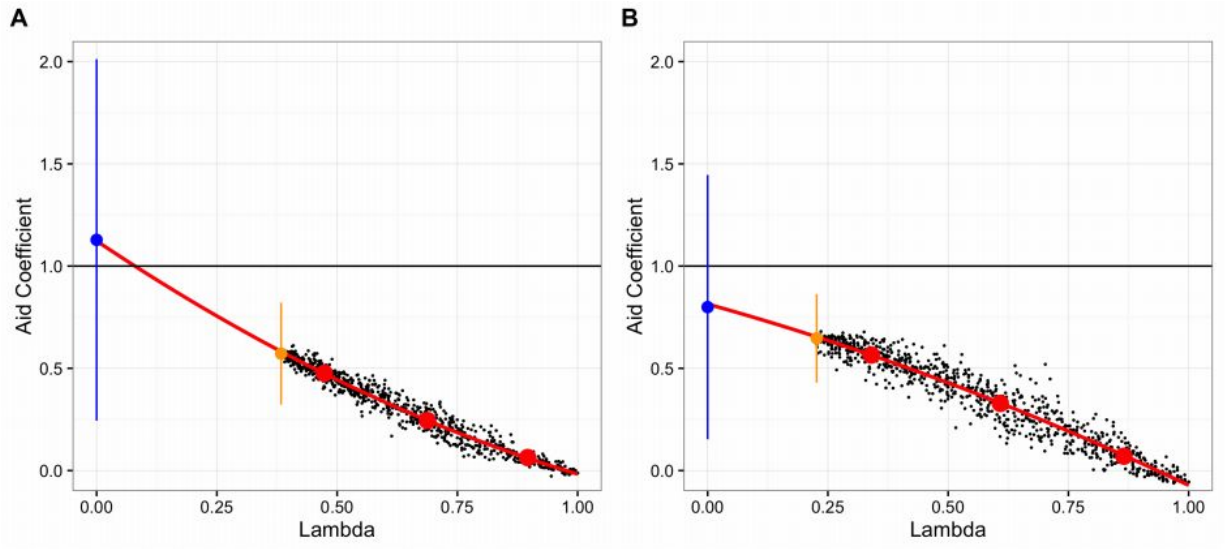


Figure 4. Panel A represents the extrapolation procedure in which the source data was imprecise relative to the procedure in Panel B. In both cases, the true coefficient (1) is captured by the geoSIMEX procedure, but not traditional linear models (orange). In the case of greater imprecision in the source data, the estimated standard errors (95% confidence) are also larger.

This process was repeated 10,000 times, and each time a random dataset was generated with different spatial characteristics to better understand the effectiveness of geoSIMEX relative to naive models. Table 1 shows the percentage of times geoSIMEX and naive models capture the true coefficient within a 95% confidence interval. On average, the geoSIMEX model captures the true coefficient 95% of the time, while the naive model captures the true coefficient 66% of the time. As spatial uncertainty increases the likelihood of the naive model capturing the true coefficient decreases. However, the ability of the geoSIMEX model to capture a significant relation decreases as spatial uncertainty grows, indicating that greater spatial uncertainty weakens evidence to reject the null hypothesis of the model coefficient being zero.

Spatial Uncertainty of Naive Model	Number of Simulations	Results Category	geoSIMEX Model	Naive Model
$0 < \lambda < 0.3$	7512	Contains True Coef.	93.1%	67.6%
		Contains True Coef. & Significant	51.5%	67.5%
$0.3 < \lambda < 0.7$	3359	Contains True Coef.	97.9%	61.7%
		Contains True Coef. & Significant	3.6%	61.1%
$0.7 < \lambda < 1$	28	Contains True Coef.	100%	57.1%
		Contains True Coef. & Significant	0%	42.9%
$0 < \lambda < 1$	10899	Contains True Coef.	94.6%	65.8%
		Contains True Coef. & Significant	36.6%	65.5%

Table 5. Summary statistics of simulations in which data of variable spatial precision (represented by lambda) was generated, and both a naive and geoSIMEX model were generated. “Contains True Coef” refers to the 95% confidence interval capturing the coefficient. Percentages indicate the proportion of simulations within each category.

Example of geoSIMEX using AidData

This section provides an example using geoSIMEX with data from AidData's geo(query) tool (http://geo.aiddata.org/query). Here, we examine the impact of aid allocated towards government and civil society (e.g. institutional capacity building, economic policy planning, civil service reform) on enhancing development in Uganda, using Uganda's third administrative division (sub-counties) as the unit of analysis (n=965). As a proxy for economic development, we use average nighttime lights in sub-counties. We use nighttime light data from 2000 and 2010, and included all government and civil society aid projects allocated from 2000 to 2010. Government and civil society aid includes 72 discrete projects from 30 different donors allocated across 389 project locations, totaling in US$9.6 billion committed. As anticipated, the underlying data contains spatial imprecision; across the 389 project locations, 97 are coded as precision code 1 (sub-county level spatial certainty), 13 as precision code 2 (county-level spatial certainty), 240 as precision code 3 (district-level spatial certainty), and 39 as precision code 6 or 8 (country-level spatial certainty), yielding a spatial uncertainty g value of 0.12.

In this illustrative example, we estimate the following model:

$$\textbf{\textit{Nighttime Lights Change}} = \beta 0 + \beta 1 * \textbf{\textit{log}}(\textbf{\textit{Aid}}) + \beta 2 * \textbf{\textit{log}}(\textbf{\textit{Population}}2000) \qquad \text{eq. 3}$$

where *Nighttime Lights Change* is the change in nighttime lights from 2000 to 2010, standardized in units of its standard deviation. Aid is the expected value of aid allocated to sub-counties in Uganda from 2000 to 2010 and *Population 2000* is the population of the sub-counties in 2000. Aid and population are entered into the equation in log-form due to skewness in the variables. Table 6 shows three models, including two naive models derived from equation 3: one that includes all government and civil society aid projects, and the other only includes projects with strong project documentation; that is, projects with precision levels indicating that it was allocated to a specific location (i.e., precision code 1 projects). Both naive models could be interpreted as providing evidence that aid is harming welfare; specifically, that a one percent increase in aid is associated with a 0.05 to 0.06 standard deviation decrease in nighttime lights. In both models the negative coefficient on aid is significant at the 1% level. The third model represents the results when the geoSIMEX model is employed. The coefficient on aid in the geoSIMEX model is still negative; however, the coefficient is not statistically significant. The standard error in the geoSIMEX model is over six times the size of the standard errors in the naive models, reflecting the large degree of spatial uncertainty in where government and civil society aid was allocated that is not accounted for in the naive model. Figure 5 shows geoSIMEX plots of the aid coefficient, which highlights the larger confidence interval on the aid coefficient in the geoSIMEX model compared to the naive model.

	Nighttime Lights Change		
	Naive	Naive (PC1)	geoSIMEX
	(1)	(2)	(3)
Aid	−0.062***	−0.054***	−0.047
	(0.005)	(0.006)	(0.038)
Population (2000)	−0.249***	−0.191***	−0.209***
	(0.026)	(0.027)	(0.051)
Constant	1.447***	1.039***	1.197***
	(0.135)	(0.138)	(0.359)
Observations	965	965	965
Note:			*p<0.1; **p<0.05; ***p<0.01
			Standard errors in parentheses

Table 6. Example analysis examining the impact of aid on Nighttime Lights

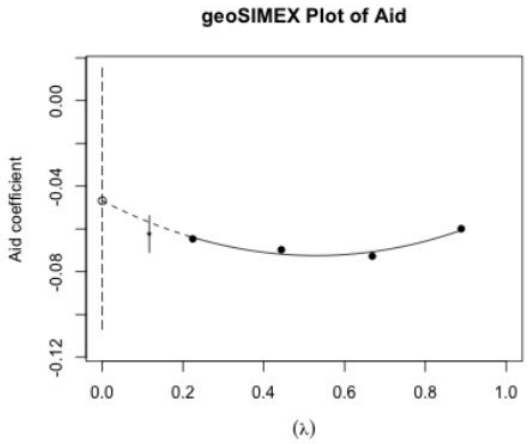


Figure 5. Beta coefficient estimates and standard errors from the geoSIMEX method (output by the R package).

As this example illustrates, accounting for spatial imprecision can fundamentally change the conclusions of an analysis - in this case, indicating that insufficient evidence exists to reject a null hypothesis. Because the geoSIMEX procedure can trivially be extended to incorporate propensity matching (i.e., by matching data and then running geoSIMEX only on that matched data), it is an important step to take to improve the quality of quasi-observational causal studies in which the imprecision of underlying data is known.