# geoSIMEX: A Generalized Approach To Modeling Spatial Imprecision

Daniel Runfola[1], Rob Marty[1], Seth Goodman[1] and Michael LeFew[1]

[1]Institute for the Theory and Practice of International Relations,
AidData, The College of William and Mary

July 19, 2016

**Corresponding Author**:
Dr. Daniel Runfola
Institute for the Theory and Practice of International Relations, *AidData*
427 Scotland Street, Williamsburg, VA 23185
Email: dsmillerrunfol@wm.edu Telephone: 508.316.9109 Fax: 757.221.4650

# 1    Abstract

Spatial imprecision exists in nearly all spatial information - from census data to satellite information, the exact geographic location to which a measurement can be attributed is rarely known. Following this, there is a large and growing set of literature examining how different classes of models can integrate information on spatial imprecision in order to more accurately reflect available data. Here, we present a flexible approach - geoSIMEX - which can provide parameter and error estimates for both linear and non-linear models while adjusting for spatial imprecision. Further, this approach can provide an estimate of the potential value of additional spatial information by distinguishing between traditional model error and additional errors introduced by imprecision. We illustrate this approach through a case study leveraging a novel, publically available dataset recording the location of Chinese aid in Africa at varying levels of precision. Using a difference-in-difference model, we integrate this information with satellite derived data on vegetation (NDVI) and a number of other spatially explicit covariates to examine if there is evidence Chinese aid has caused an increase or decrease in vegetation. Following commonly employed procedures which do not incorporate spatial imprecision, we found that Chinese Aid had a negative impact on vegetation; once spatial imprecision was incorporated into our estimates through the geoSIMEX procedure no evidence of impact was found. We use these findings to argue for the importance of incorporating spatial imprecision into analyses, and introduce new software in the R programming language to enable similar analyses for other purposes.

## 2 Introduction

The lack of exact geographic information on where measurements are obtained presents a barrier to research. This has become increasingly evident as more scholars integrate geographic data from multiple sources - for example, census, satellite, and GPS sources - to try and establish causal or predictive relationships (c.f., EXAMPLES). Many scholars have engaged in research seeking to integrate information on the precision of geographic data to improve the accuracy of modeling efforts, variably referred to as spatial (im)precision (CITE), spatial uncertainty (CITE), and spatial allocation (CITE). This paper presents a generalizeable approach to integrating information on the precision of geographic data into both linear and non-linear models - geoSIMEX. We illustrate the capability of geoSIMEX to provide more accurate parameter estimates than traditional approaches through a simulation framework. Using a novel dataset, we then apply both tradional models and a geoSIMEX model to examine the causal impact of Chinese Aid on vegetation in Africa. We use this case study to illustrate the importance of including information on spatial imprecision into analyses. Finally, we provide all data and an accompanying R package to users seeking to perform similar styles of analysis.

### 2.1 Literature Review

The literature has shown that uncertainty in the locations of where measurements were taken can produce biased estimates in empirical analyses.Perez-Heydrich et al. 2013; Rettie and McLoughlin 1999 For example,

Perez-Heydrich et al. show that regression coefficients will be biased when using raster data in conjunction with point data, where the true locations of the point data are only known to exist within some 5-10km radii of the measured location.Perez-Heydrich et al. 2013 One recommended solution is to take average raster values within a buffer encompassing where the point could have fallen, instead of the single raster value associated with the point.Perez-Heydrich et al. 2013; Rettie and McLoughlin 1999

Another practice to address spatial uncertainty is to aggregate to some higher spatial scale where there is no spatial uncertainty.Perez-Heydrich et al. 2013 However, aggregating to a higher spatial scales can introduce bias. A large body of literature shows that analyzing data at different levels of aggregation can produce different regression coefficients and correlation coefficients.Clark and Avery 1976; SELVIN 1958; Gotway and Young 2002; Gehlke and Biehl 1934; Cramer 1964 Estimates from aggregate data are invalid for disaggregate data due to different magnitude of correlations and regression coefficients. For example, a seminal study by Robinson illustrates this point. Using data from the 1930 US census, Robinson found a negative correlation between the proportion of immigrants in a state and average literacy levels, but a positive correlation between being an immigrant and literacy level.Robinson 2009 A number of techniques have been proposed to address aggregation bias. Approaches focus on dis-aggregating data to finer scales; however, these involve using covariats to help predict the variable of interest.Gotway and Young 2002; Zhu et al. 2004

## 2.2  (geo)SIMEX

The simulation and extrapolation method (SIMEX) provides one solution to address measurement error in covariates (CITE). SIMEX exploits the fact that greater measurement error will lead to greater 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.

SIMEX simulates error using the following equation:

$$X(\lambda) = X + \sqrt{\lambda}U \tag{1}$$

where X is the measured covariate, U is the variance of the measurement error and $\lambda$ is a parameter that simulates additional measurement error. $\lambda = 0$ corresponds to the original amount of measurement error in the measured covariate, X. SIMEX uses increasing values of $\lambda$ (e.g., 0.5, 1, 1.5, and 2) to estimate models with simulated amounts of additional measurement error. SIMEX then estimates a trend between $\lambda$ and the coefficient on X, and uses the trend to extrapolate back to $\lambda = -1$ (point of no measurement error).

In geoSIMEX, we take this approach and modify $\lambda$ to capture the degree of spatial imprecision in a given dataset, relative to the maximum amount of imprecision that could be observed. The practical need for a geographic version of SIMEX can be demonstrated through a hypothetical example, in which a researcher seeks to measure the degree to which the amount of precipitation across 200 fields can explain the number of apples grown in

each field. These hypothetical researchers measure precipitation in each of 200 fields using satellite imagery, and counts each of the apples by hand. In the simplest form, these researchers then use these measurements to model the relationship between rainfall and apples following:

$$\text{CountOfApples} = \beta_0 + \theta * \text{Rainfall} + \sum \beta_j * X_j + \epsilon \tag{2}$$

in which $\beta_j$ and $X_j$ represent vectors of parameter estimates and controls, respectively. In this equation, a common goal would be to estimate $\theta$ with a high degree of accuracy, as this parameter helps to explain the degree to which precipitation impacted the count of apples.

In this example, if each field was a perfect square of the same resolution as the satellite imagery (and had perfect overlap), then the measurements of rainfall would have no measurement error due to imprecision. In practice, this rarely is the case, and thus the parameter $\theta$ can be biased due to imprecision in measurements. Consider, for example, the worst case of imprecision in which a coarse reolution satellite sensor provides only one measurement of rainfall for every crop. A responsible researcher seeking to use this data might leverage a model averaging framework to overcome concerns of the ecological fallacy - i.e., randomly distributing rainfall to each crop, fitting equation 2, and repeating this hundreds of times to estimate the range of possible coefficients that could be observed given the available data. However, in cases of complete spatial imprecision, these estimates will be heavily biased towards 0. This has two implications - first, the mean estimated coefficient will likely under-estimate the true coefficient, and second

the distribution of possible coefficients (i.e., at a 95% confidence level) will be centered at this biased coefficient. This suggests that not only can mean estimates be incorrect, but under cases of spatial imprecision statements regarding the significance of variables can also be biased.

While it is rare that satellite measurements perfectly line up with units of analysis, it is also rare that researchers only have one measurement for all units of analysis. More reasonably, researchers commonly have a mixture - i.e., satellite imagery that may cover portions of some crops but not others, leading to variable imprecision across the study area. geoSIMEX takes advantage of the known level of precision in the data available ot the research (which we define as $\lambda$), and intentionally adds imprecision to establish the relationship between additional imprecision and coefficient bias. The geoSIMEX procedure then uses this relationship to back-extrapolate to an unbiased coefficient estimate, i.e. the value at $\lambda = 0$. Finally, we employ a bootstrapping procedure to estimate the additional variance attributable to spatial imprecision.

# 3 Data and Methods

This study leverages a novel dataset on the location of Chinese international aid in Africa available at varying levels of precision (i.e., the exact location of each aid project is not always known). It integrates this information with a variety of other ancillary datasets, including the NASA Long Term Data Record (LTDR) to examine the causal impact of Chinese aid on vegetation. Employing geoSIMEX, we use this case study to illustrate the importance of incorporating information on spatial imprecision into analyses.

## 3.1 Data

To conduct this analysis, a new dataset on the location of Chinese aid is derived through a methodology designed to Track Underreported Financial Flows (the TUFF methodology). This dataset is merged with a number of existing datasets, retrieved from publically available sources as described below.

### 3.1.1 Study Area and Scope

China has increasingly taken a larger role near home in providing aid to developing countries. In this analysis, we examine the impact of Chinese aid in southeast asia, explicitly focusing on Cambodia, Laos, Myanmar, Thailand and Vietman. Using the Global Administrative Areas (GADM; CITE) database, we aggregate Chinese aid, covariate and outcome information to the second-level administrative units within each country (see figure 3).

### 3.1.2  Aid Information - Tracking Underreported Financial Flows (TUFF)

Unlike many donors, China does not publically report the international aid they send (CITE). (Placeholder for TUFF details) Due to the considerable spatial imprecision generated by the TUFF process, we use this variable as an illustrative case of how the geoSIMEX procedure can be used to provide better estimates than traditional approaches.

### 3.1.3  Ancillary Datasets

Covariate data is collected from a variety of sources, summarized in table 1. Our outcome measure - fluctuation in NDVI - is derived from the NASA Long Term Data Record (LTDR) dataset. While relatively coarse resolution, this dataset represents the longest consistent record of NDVI available at the global scale. To facilitate our difference-in-difference modeling efforts, we further select a number of covariates we believe could also impact shifts in NDVI (other than Chinese aid). These include:

1. Long-term climate data from the University of Delaware, providing precipitation and temperature data at a monthly time-step for the full data record, which is permuted to produce yearly mean, minimum, and maximum values for each project location.

2. Population Data is retrieved from CIESIN at Columbia University, specifically leveraging the Gridded Population of the World (GPW) data record.

3. Slope and Elevation data are derived from the Shuttle Radar Topography Mission (SRTM).

4. Distance to rivers is calculated based on the USGS Hydrosheds database.

5. Distance to roads is calculated based on the Global Roads Open Access Dataset (gRoads), which represents roads circa 2010, though the actual date of datasets is highly variable by country.

6. Urban travel time, calculated by the European Commission Joint Research Centre.

7. Nighttime Lights are retrieved from the NOAA Earth Observation Group, calculated from the Department of Defense Defense Meteorological Satellite Program (DMSP). Lights values are temporally intercalibrated following the procedure outlined in Weng 2014

Each of these datasets are processed and aggregated according to their average values within each district included in this analysis. Further, the size of districts are controlled for to mitigate the challenge of variably-sized districts across the study area. In cases where covariates were measured at a resolution coarser than the unit of observation, the relative area of overlap was used to generate a weighted mean. Further information on this decision can be found in the discussion.

## 3.2 Methods

Two different modeling processes are followed. First, we use a monte carlo simulation procedure to illustrate the relative accuracy of geoSIMEX as

contrasted to other procedures. Second, we apply geoSIMEX to the case of Chinese Aid in Africa.

### 3.2.1   geoSIMEX

We employ geoSIMEX to identify the impact of Chinese Aid on vegetation, measured by a satellite-derived measure of NDVI. We specifically seek to adjust for bias in the relative spatial imprecision of where Chinese Aid is located - imprecision largely due to the lack of formal information published by the Chinese government. We describe this process through an illustrative example, in which we attempt to solve the following simplified equation:

$$NDVI = \theta * \text{Chinese aid} + \epsilon \tag{3}$$

where, for the sake of example, Chinese aid is defined to be causally and positively related to a measure of NDVI by a one-to-one relation (e.g., $\theta = 1$), and $\epsilon$ is a random error term. Aid is measured with spatial imprecision, and through using geoSIMEX we account for the spatial uncertainty to accurately estimate the model coefficient, $\theta$.

In figure **??**, we present a hypothetical country with sixteen districts for which we seek to solve equation 3. Four of these sixteen units of analysis, districts 5,6,7, and 8 are distinguished on the map. Within this study area, a hypothetical data set contains three geocoded Chinese aid project locations of various levels of spatial precision, projects A, B, and C. Project A is assigned a coordinate pair in District 5 and had strong documentation, resulting in precise geographic information (i.e. an exact latitude and lon-

gitude). Due to weaker project documentation, location B has a precision level indicating that it was allocated anywhere in the region that includes districts 5,6,7, and 8. Project C has very uncertain spatial information, such that it may be anywhere in the country. The area (in square kilometers) of the unit of analysis in which each project location may have been allocated gives us the size of the project locationâĂŹs area of coverage, which is summarized in table 2. Using the spatial overlap between each region of interest (the sixteen districts) and the area an international aid project might exist in, we calculate a probability that each district contains a given project:

$$V_t = \sum_{S}^{s=1} U_s \left( \frac{a_{st}}{\sum_{T}^{t=1} a_{st}} \right) \tag{4}$$

This represents the "no information" case - i.e., probabilities are only based on geographic overlap, as opposed to integrating other factors which might mediate where aid is allocated. This equation can be modified to incorporate more information on spatial location, thus allowing a researcher to trade off additional assumptions or information about factors that mediate spatial allocation in exchange for higher degrees of spatial precision (i.e., through dasymetric mapping approaches).

We use these probabilities in a modified version of the SIMEX procedure, geoSIMEX. In traditional SIMEX, measurement error is captured through scaling an estimated distribution of errors. In our context, greater spatial imprecision is reflected by increasing the area of coverage of a project, thus expanding the probability a project could be located in a wider set of units of analysis. geoSIMEX exploits the fact that greater spatial imprecision will

13

bias results towards 0 - i.e., if all data was measured at the country scale and every district had an equal probability of receiving aid, any results found using districts as the unit of analysis would be the equivalent of random noise. geoSIMEX (a) calculated the initial level of spatial imprecision in a given dataset, and then (b) simulates additional spatial imprecision to establish this relationship between spatial imprecision and covariate bias. Based on this relationship, it extrapolates to a point with zero spatial imprecision, thus providing an estimate of the unbiased coefficient.

**The first step** of geoSIMEX is to calculate the initial level of spatial imprecision in the given dataset, defined by $\lambda$. To reflect imprecision across a given set of international aid projects and a set of units of analysis (i.e., districts), we calculate ($\lambda$) following:

$$\lambda = \frac{\sum_i^P Area\ of\ Coverage_i}{\sum_i^P Total\ Possible\ Area\ of\ Coverage_i} \tag{5}$$

where $i$ is an individual project out of $P$ total Chinese Aid projects. $Area\ of\ Coverage_i$ is project $i$'s known area of coverage defined by the available documentation - i.e., the geographic area across which a project could be located. $Total\ Possible\ Area\ of\ Coverage_i$ is the area of coverage of project $i$ under complete spatial imprecision - e.g., the geographic area of the study area.

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

14

of different levels of precision result in *lambda* values between these two extremes, providing a single linear measurement of spatial imprecision.

**The second step** involves estimating a naive model (which can be of variable functional forms; for illustration we use ordinary least squares regression), where imprecision in aid is ignored. For each unit of observation, the value of aid used for modeling is the expected value of aid calculated following the geographic overlap-based probabilities described in (EQUATION!!). In figure 2, we provide an example of the geoSIMEX procedure is applied to a dataset with an initial $\lambda$ value of 0.4, indicating the set of data provided was of moderate spatial precision relative to the units of observation. In this figure, the x-axis represents the $\lambda$ value (spatial imprecision) for a dataset, with higher values indicating more imprecision. The y-axis represents the estimated value of $\theta$ in equation **??**. In 2a, the orange line represents the 95% confidence interval of the coefficient on aid in this naive model. The horizontal black line represents the true model coefficient ($\theta = 1$), which the naive model fails to capture.

In **the third step**, additional imprecision is simulated by randomly decreasing the precision of observations (in this case, Chinese Aid projects). For example, a project that has a measurement with an exact latitude and longitude will randomly be assigned a lower level of precision - i.e., a county, state, or even the entire country. Using these new, reduced levels of precision a model is fit in an identical fashion to step 1, and the estimated $\theta$ parameter, standard errors of the model,ad $\lambda$ value for a given permutation are saved. This process is repeated 10000 times. In figure 2b, the black points represent individual iterations, with the saved model coefficients (y axis) and their

associated $\lambda$ (x axis) values.

**The fourth step** subdivides this set of iterations into three equally-sized bins based on the level of spatial uncertainty ($\lambda$) of the aid variable (e.g., if $\lambda$ 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 $\lambda$ values are calculated within each bin, represented as red dots in figure 2c. A quadratic trend is fit on the resulting average coefficient and lambda values. The trend is then extrapolated back to $\lambda = 0$, thus providing an estimate of $\theta$ with perfect spatial precision. In figure 2d, the red line represents the extrapolated trend, and the blue dot represents the extrapolated estimate of the coefficient on aid.

**In the fifth step**, the variance and standard errors of these estimates are calculated. We employ a bootstrapping method to calculate the component of the standard error resulting from spatial imprecision. Here, a point from each bin is sampled, a quadratic trend is fit on the resulting values, and the trend is extrapolated back to $\lambda = 0$. This process is repeated 10000 times (defined as $R$). In figure 2e, each blue line represents one extrapolated trend and $\lambda = 0$ estimate. We use a variance equation originally developed to capture model selection uncertainty to seperately incorporate both original standard errors and the additional error from spatial imprecision (CITE Burnham and Anderson):

$$var(\hat{\bar{\beta}}) = \sum_i^R \frac{1}{R}\{var(\hat{\beta}_i) + (\hat{\beta}_i - \hat{\bar{\beta}}_i)^2\} \tag{6}$$

where $R$ is the number of extrapolated coefficients (in this example, 1000).

$var(\hat{\beta}_i)$ is the standard error of each extrapolated coefficient, calculated by fitting a quadratic trend on the standard error estimates from each bin extrapolating back to $\lambda = 0$ and collecting the resulting standard error value. $(\hat{\beta}_i - \bar{\hat{\beta}}_i)^2$ captures the component of the variance from spatial uncertainty, where each extrapolated coefficient estimate is subtracted by the mean of all extrapolated coefficient estimates.

### 3.2.2  Simulations

We employ a monte carlo procedure to determine the accuracy of geoSIMEX compared to other approaches to handle spatial imprecission. The procedure generates data to model the relation between Chinese aid and NDVI according to equation 3. Each simulation procedes in the following 7 steps:

1. One of three hypothetical countries with different administrative heirarchices is generated: (1) a country with 60 subcounties, 30 counties, 10 districts, and 2 regions, (2) a country with 120 subcounties, 40 counties, 20 districts, and 5 regions, or (3) a country with 120 subcounties, 60 counties, 30 districts, and 10 regions. Each subcounty is randomly assigned a spatial area and a probability of recieving aid.

2. 50 to 250 aid projects are randomly allocated to subcounties, according to the assigned probability of a subcounty receiving aid.

3. NDVI is generating according to equation 3. Specifically, NDVI is defined as the number of aid projects plus random error.

4. Each aid project is given a code indicating the spatial precision that a

17

researcher might see. Codes range from indicating the project fell within a sub-county (no spatial imprecision), to falling somewhere within a county, district, region, or the country.

5. The expected value of aid is calculated for each subcounty. For projects with spatial imprecission, aid is disbursed to subcounties based on spatial area, with larger subcounties having a larger probability of recieivng aid. For example, consider a project where $1 million was disbursed to either subcounty A (spatial area of 400km) or subcounty B (spatial area 100km). Because subcounty A is four times as large as subcounty B, subcounty A will be assigned four times as much aid from the project as subcounty B. Specifically, subcounty A will be assigned $800,000 and subcounty B will be assigned $200,000. Disbursing aid using spatial area assumes that each location (e.g., each sq. kilometer) within a country has an equal chance of receiving aid.

6. Equation 3 is estimated using two naive models that do not incorporate spatial imprecision: the first uses the expected value of aid, and the second only includes aid projects with complete spatial preceision.

7. Equation 3 is estimated using geoSIMEX and a model averaging approach. In the model averaging approach, subcounties are assigned a random probability of receiving aid and the expected value of aid is calculated according to that probability. 500 models are estimated, where the average coefficient is used with the standard error estimated using equation 6.

This process is repeated 100,000 times.

### 3.2.3 Chinese Aid in Africa

In an illustrative case study of geoSIMEX, we examine the causal impact that Chinese aid distributed from 2005 to 2010 had on vegetation (measured using NDVI) in southeast asia.

$$Y_i = \beta_0 + \theta * Aid_i + \sum_{i=k}^{n}(\beta_k * X_i) \tag{7}$$

Where $Y_i$ is the difference in the average forest loss post-2011 and pre-2004 in district $i$; $Aid_i$ is the estimated aid that went to each district; $X_i$ is a vector of control variables; $\beta_k$ is a vector of parameters for each beta covariate, and $\theta$ is the estimated impact of aid.

Aid is the log of total commitments to administrative divisions allocated from 2005 to 2010.

Controls include aid temperature (maximum, minimum, and average), precipitation (maximum, minimum, and average), lnyx, and nighttime lights. Variables are averaged within administrative units, then averaged into pre-aid and post-aid groupings (pre-aid includes average within 2001 to 2004 and post-aid includes average within 2011 to 2013).

19

# 4 Results

### 4.0.4 Simulations

Table (NUM) shows results from the simulation analysis. Results are subset according to the spatial imprecession of the Chinese aid data. The naive model using the expected value of aid captures true coefficient about 50% of the time within a 95% confidence interval when there is low spatial uncertainty ($\lambda < 0.3$), and the ability for the model to capture the true coefficient declines as spatial imprecesion increases. The naive model using the subset of aid projects measured with complete spatial uncertainty captures the true coefficient 60% to 70% of the time depending on the level of spatial imprecesion in the data. The model averaging approach performs well under low levels of spatial imprecession; however, as spatial imprecession increases the ability for the model averaging approach to capture the true coefficient decreases. Under low levels of spatial imprecession the model averaging approach captures the true coefficient 85% of the time, but under high levels of spatial imprecession the model captures the true coefficient 46% of the time.

The geoSIMEX model outperforms the model averaging approach in its ability to capture the true coefficient at all levels of spatial imprecession. At low levels of spatial imprecession the geoSIMEX model captures the true coefficient 90% of the time; however, it captures both the true relation and statistical significance 52% of the time compared to 84% of the time from using model averaging. Higher spatial imprecesion leads to larger standard errors in the geoSIMEX model, reflected by the percentage of time

the geoSIMEX model captures the true relation and statistical significance declining as spatial imprecesion grows. At high levels of spatial imprecesion ($\lambda > 0.7$), the geoSIMEX model captures the true relation 100% of the time; however, large standard errors results in it capturing the true relation and statistical significance 0% of the time.

### 4.0.5   Chinese Aid in Africa

# 5  Discussion

Broadly, this paper is related to papers that seek to overcome issues related to the Ecological Fallacy - i.e., the fallacy of measuring data at a course resolution and applying it to units of a finer resolution. By treating this "fallacy" as a source of explicit uncertainty in the modeling process, we argue that the use of relatively coarse resolution data not only can provide more insight into processes of interest, but that by ignoring coarse-scale information researchers may be commiting a seperate fallacy - that of ignoring "known unknowns". This can be particularly troublesome in analyses which seek to establish causal relationships between variables, as bias that results from spatial imprecision can lead to incorrect statements of the statistical significance of such relationships.

While geoSIMEX mitigates many issues related to spatial imprecision, it is not a silver bullet solution. First and foremost, the ability of geoSIMEX to uncover the true coefficient is heavily influenced by the initial precision of the observed data. If a researcher attempts to use data which has complete imprecision (i.e., $\lambda = 1$), the results provided by geoSIMEX will be meaningless as no information is available from which a trend can be estimated. As $\lambda$ approaches 0 (perfect precision relative to the units of observation), geoSIMEX will provide increasingly accurate estimates as it has an increased number of $\lambda$ units to derive data for - i.e., there are more variable simulations across observations of $\lambda$. This feature of geoSIMEX is novel when compared to traditional SIMEX. In the traditional SIMEX process, an infinite amount of error can be simulated in any given variable by increasing the width of

the error distribution. In geoSIMEX, the maximum amount of imprecision in the system is limited by the spatial configuration of the study area (i.e., the maxmium imprecision is observed when all measurements are taken at the scale of the entire study area).

There are many directions for future work. First, the process described here should be generalizeable to multiple dimensions of spatial imprecision - i.e., imprecision in multiple attributes of the data. Such multidimensional approaches to SIMEX have been examined in other fields (CITE), but applying them to the case of geographic imprecision should help to overcome many issues surrounding data integration alluded to in this piece. Second, the authors hypothesize that the SIMEX procedure could be applied not only to cases of imprecision in observed variables (the ecological fallacy), but may also provide a solution to the ever-present concern of arbitrary units of observation (the modifiable area unit problem, or MAUP). By treating the boundaries of units of observation as a unique cause of spatial imprecision (i.e., the exact - or correct - boundaries may not be known), geoSIMEX could be modified to provide estimates unbiased by covariates or boundary selection. However, both computational and methodological barriers remain to this solution. Finally, in non-hierarchical datasets the application of geoSIMEX is currently not feasible due to a lack of a relevant "weights" matrix to establish how data should be permuted at higher levels of $\lambda$. For example, in a raster dataset coarsening data can be done simply by aggregating nearby grid cells; conversely, if one sought to examine the impact of global international aid on non-contigious protected areas, it is unclear how imprecisions should be introduced into the system.

23

geoSIMEX has a secondary benefits in the ability to seperately identify traditional model errors and errors introduced due to spatial imprecision (see eq. 6). By breaking these sources of error out, one could provide estimates of how much value additional information on spatial imprecision could provide to a given analysis. This information could be useful in the justification of more precise approaches to spatial measurement.

# 6 Conclusion

In this piece, we introduce a flexible method to account for spatial imprecison in datasets for the purpose of fitting linear or non-linear models - geoSIMEX. We illustrated the relative accuracy of geoSIMEX as contrasted to other methods which do not account for spatial imprecision, finding that geoSIMEX outperforms all other tested approaches under cases of imprecision. Additionally, as an illustrative example we applied geoSIMEX to the case of Chinese Aid's impact on NDVI in southeast Asia. In models which did not account for spatial imprecision, it was found that Chinese Aid had a negative impact on forest cover. By accounting for spatial imprecision, we illustrated that in fact there was insufficient evidence to draw this conclusion.

We identify a number of advantages of the geoSIMEX procedure.

Further, this approach can provide an estimate of the potential value of additional spatial information by distinguishing between traditional model error and additional errors introduced by imprecision. We illustrate this approach through a case study leveraging a novel, publically available dataset recording the location of Chinese aid in Africa at varying levels of precision. Using a difference-in-difference model, we integrate this information with satellite derived data on vegetation (NDVI) and a number of other spatially explicit covariates to examine if there is evidence Chinese aid has caused an increase or decrease in vegetation. Following commonly employed procedures which do not incorporate spatial imprecision, we found that Chinese Aid had a negative impact on vegetation; once spatial imprecision was incor-

porated into our estimates through the geoSIMEX procedure no evidence of impact was found. We use these findings to argue for the importance of incorporating spatial imprecision into analyses, and introduce new software in the R programming language to enable similar analyses for other purposes.

# 7 Tables

| Data Sources | |
|---|---|
| **Data Name** | **Source** |
| Chinese Aid Locations | AidData[1] |
| Gridded Population of the World | Center for International Earth Science Information Network[2] |
| Nighttime Lights | Defense Meteorological Satellite Program[3] |
| Precipitation and Temperature | University of Delaware (Willmott and Matsuura 2001)[4] |
| Urban Travel Time | European Commission Joint Research Centre[5] |
| Distance to Rivers | World Wildlife Fund [6] |
| Vegetation | NASA LTDR[7] |
| Distance to Roads | CIESIN gRoads[8] |

Table 1: Data sources used in this analysis.

---

[1] http://china.aiddata.org

[2] http://sedac.ciesin.columbia.edu/data/collection/gpw-v3/sets/browse

[3] Stable Lights retrieved from http://ngdc.noaa.gov/eog/dmsp.html

[4] Variables derived from these product included the average precipitation (P) and temperature (T) before a project was implemented (from 1992), the linear trend in P and T from 1992 to the project implementation, the average temperature from the date the project was implemented until the end of the temporal record(2012), and the post-project trend through 2012. Absolute measurements of each variable were also retained.

[5] http://forobs.jrc.ec.europa.eu/products/gam/download.php

[6] http://hydrosheds.cr.usgs.gov/index.php

[7] http://ltdr.nascom.nasa.gov/cgi-bin/ltdr/ltdrPage.cgi

[8] http://sedac.ciesin.columbia.edu/data/set/groads-global-roads-open-access-v1

| Project Location | Relative Precision | Known Aid Location (Geographic Size) |
|---|---|---|
| A | Very High | Populated Area in District 5 (3 km$^2$) |
| B | Moderate | Districts 5-8 (32,000 km$^2$) |
| C | Very Low | Entire Country (112,500 km$^2$) |

Table 2: Example of the geographic area across which aid projects might be located given limited spatial information.

Table 3: Simulation Results

| Spatial Uncertainty of Naive Model | Number of Simulations | Results Category | Naive Model | Naive Model (Subset) | Model Average | geoSIMEX Model |
|---|---|---|---|---|---|---|
| $0 < \lambda < 0.3$ | 1262 | Contains True Coef. | 52.9% | 62.6% | 85.2% | 86.9% |
| | | Contains True Coef. & Sig. | 52.9% | 61.1% | 84.9% | 69.7% |
| $0.3 < \lambda < 0.7$ | 914 | Contains True Coef. | 44.4% | 63.2% | 75.9% | 93.4% |
| | | Contains True Coef. & Sig. | 44.2% | 57.5% | 66.5% | 25.9% |
| $0.7 < \lambda < 1$ | 678 | Contains True Coef. | 23.9% | 72.4% | 47.5% | 100% |
| | | Contains True Coef. & Sig. | 23% | 61.8% | 17% | 0% |
| $0 < \lambda < 1$ | 2854 | Contains True Coef. | 43.3% | 65.1% | 73.3% | 92.1% |
| | | Contains True Coef. & Sig. | 43% | 60.1% | 62.9% | 39.1% |

"Contains True Coef" refers to the 95% confidence interval capturing the coefficient.

# 8 Figures
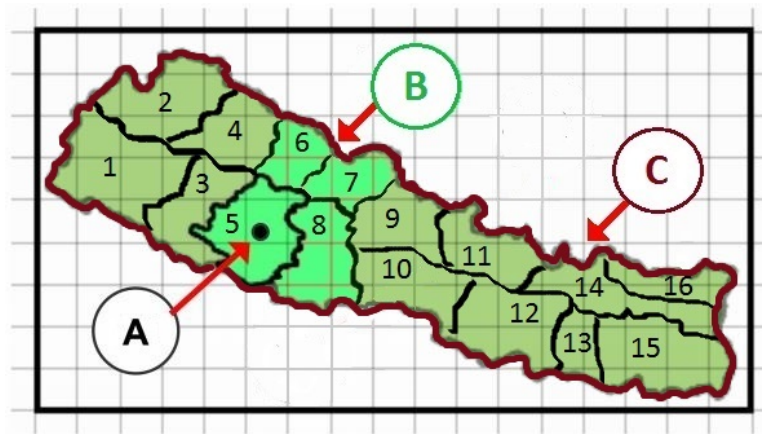


Figure 1: Hypothetical example of spatial imprecision in aid allocation.
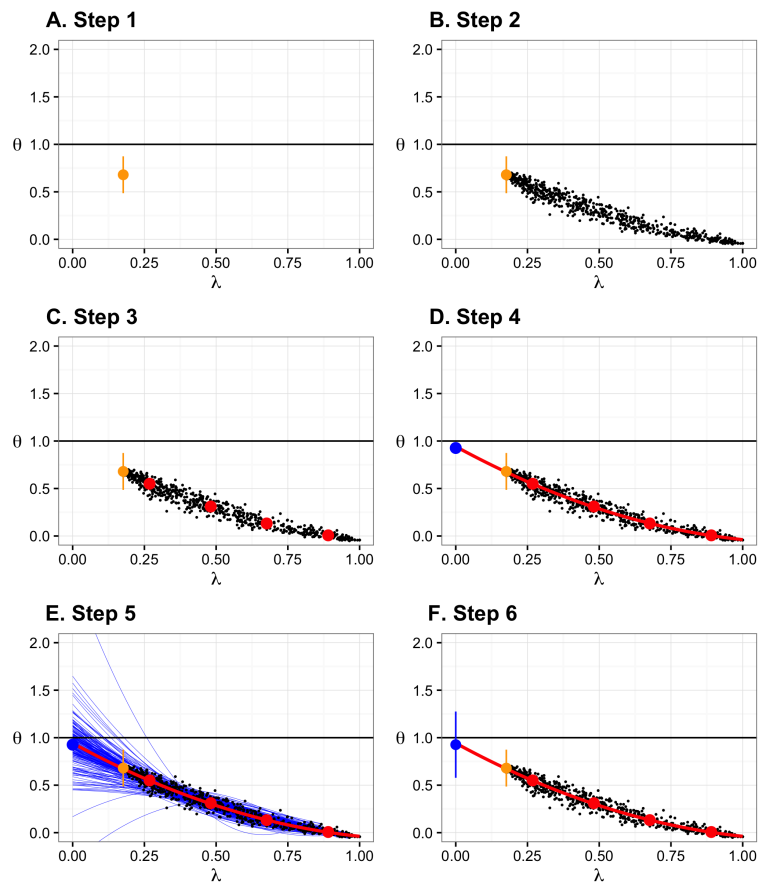
Figure 2: Steps of the geoSIMEX procedure.

Figure 3: Study area and units of observation.

31
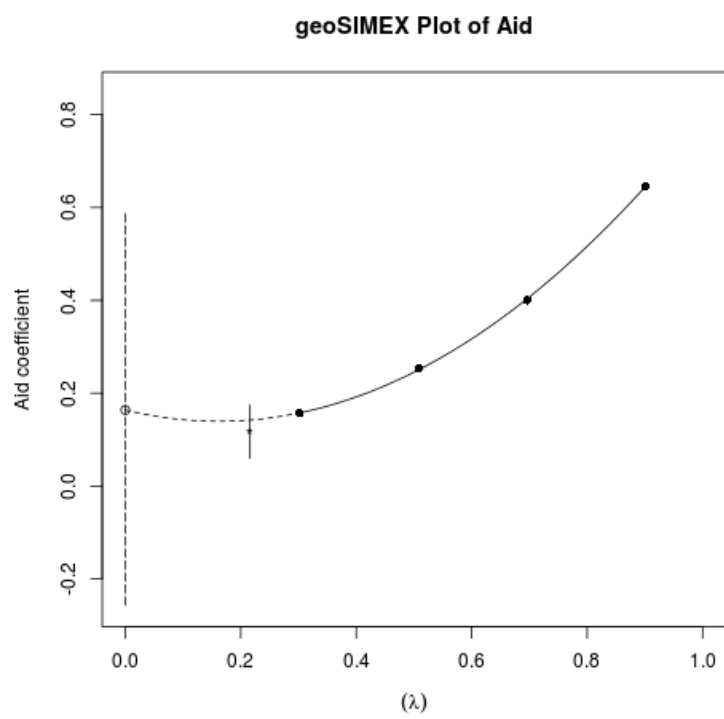
Figure 4: geoSIMEX Plot of Aid.

# 9 Acknowledgements

# References

Clark, W. A. V. and Karen L. Avery (1976). "The Effects of Data Aggregation in Statistical Analysis". In: *Geographical Analysis* 8.4, pp. 428–438. ISSN: 0016-7363. DOI: 10.1111/j.1538-4632.1976.tb00549.x.
URL: http://wm.summon.serialssolutions.com/2.0.0/link/0/
eLvHCXMw3V3PS8MwFA7Ti4KIP3E6IScvo6Vp0qQ9eChT8ebBeS5Jk8wdNkUnuv_
epE3abjIEj15LeKR5X9-P8r0vAOA4jIK1mEAjnfKEC0wSmzC1UohpqoRGMZe0lGujVv7m0vbZf3H8bcvSu
eNUSbpVqrsZ_fnnAs_oqah54TAPW5Jsbs5m2cyXuWkGWQ_YMdow05oYiWjAsAs7qg2LhOKVuJl24EE6MZC
kc1yTyHsyuPvZa2GjJhp40xtgprq7C2Cmer-LqyMuozOS0X12oePD1umR49tZO6ehg1aTtm9V3a_
vWdQq2jfW3Y5Wo1s5rMq0Q8PgD7rrWAeQ2JQ9BT8yOw4325PAbMQAM6aMAXDS00YAcacDqHHWhA7_
kTML67HY_uA3dxRlCyLAoUKTMmqKJcaWS1Xq36ayYTK9yEkIgEiwknTGcZjwjXSVRGkqZYJBJzUy8ifAr2
qOUx5BqDCWHDEsFCmAicCm-VUi1gzKVTKy6wPsD-I4rXWSSl-d0wfDPyZFe4Tei-
sAJRZalqJ8z8ZvQC7La4HYHvx9qEuQe9z9g0Os2ax.

Cramer, J. S. (1964). "Efficient Grouping, Regression and Correlation in Engel Curve Analysis". In: *Journal of the American Statistical Association* 59.305, pp. 233–250. ISSN: 0162-1459. URL: http://wm.summon.
serialssolutions.com/2.0.0/link/0/eLvHCXMwlV3LTsMwEFxVnCoh3ohHK_
kDCMQbO3ZPCFWtOKNyjhw_EIhGpbT09_EjgQoOiA9IFNnJ7Gw8OwNQ4HWe_
cAELYwZaW4cjhgV2pZaa4dUGaasEVH8vjVqBV2oYxBZRpVgPNP3dKl-tTcYugSBt4u3LIRHhUPWNknDIzH
1Ie_kLfiJDTfXjsBnOSkmQz3xJXb7k0_vPhDmCv5ZjkLr0Uh9CzzRH0A61MrszHMJ1E5whfcEj8-
eTr1xV5sE9JFNsQ1RgyDrkdSSlHnhsyCepXMl4vPyzpnExOYDadzMb3WZuokL1ISrMRK3OnkSnpiiI3BjF
e9Ch5AMpjz4cUfl19CP1iKJVXYAHZWy7UdQm8z_wSas74h.

Gehlke, C. E. and Katherine Biehl (1934). "Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material".

In: *Journal of the American Statistical Association* 29.185, pp. 169–170. ISSN: 0162-1459. URL: `http://wm.summon.serialssolutions.com/2.` `0.0/link/0/eLvHCXMwlV3NS8MwFH8MTwPxW_wavIPXatNkTXeSURxePLmdS2IzEFxXXWXgX7_` `3kg63HRRvgUBIacj7te_3ASCTuzjauRMIVWeDMnudaqWcFKyHlKxMsCYu-8K6bakVrEMdmWTpWYK-` `p09wyb67-yTRVNbOQ_ORcXgUN1nbJA26iQUVuGCT-3Mfa58-SdgmiYTq7zAPfROZHcJkrcYJ9JHlbINRvW` `M8dHcFBCyxxGE7CMXRcdQJdxpLBivkUTB7a_xgcixc4n6L_9UTVCyf1vEICg_` `jy9u14hsc5R3cEshyNnXeboCKFtEbOERkLHLPICp9N4w_yGdyOHsf5U-T3XtTByaKgLwAmghftA8hz2DdM` `XfPqj_lr6BKGVIHWdQN7zeeX60FnOVsB3VWtVQ`.

Gotway, Carol A. and Linda J. Young (2002). "Combining Incompatible Spatial Data". In: *Journal of the American Statistical Association* 97.458, pp. 632–648. ISSN: 0162-1459. DOI: `10.1198/016214502760047140`. URL: `http://wm.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwnV1bS8MwFD5MfRmI92HVwUD` `iYjDH6CITyFpEl_c1LVjf9-cNNvqBRFfSiG9pD3pyXdOvn4HgGQXSfzFJ9DEcjmUiuRDnDCtMSmz1CibZl` `1QoWqhDB2gsn6T23fi0xaX6J4RSjLMmv3t5jrCKFq62hpMYabGRZQnCUs8enlWdmvg6lQzlZnOYoT9qai4` `DmDzOSd6qj7abEauVFC5FOMh-3eNgtQcf_PscObAVcOrhuBtIudMxkD7oIRRsl5304da5D-` `XISA1ROGCMVW72YQYU77lTkmh7Aw-j2_uYuDiUW4jIrUh4TriwdGmIKmnFpqNWUUeMgAEegwOvNOdG4gKI` `VMF7y0CdOyyBPSgO2JVPxJ7X_ZO4cwMIQomTKijANruXIhsKJWZZZpZbgsiwjOF29dvDWSGsKHIgUXaBpn` `h66vV-CzRCazX05npQ2c-_gCDnhRs`.

Perez-Heydrich, Carolina et al. (2013). *Guidelines on the use of DHS GPS Data: DHS Spatial Analysis Reports 8*. URL: `https://dhsprogram.com/` `pubs/pdf/SAR8/SAR8.pdf`.

Rettie, WJ and PD McLoughlin (1999). "Overcoming radiotelemetry bias in habitat-selection studies". In: *CANADIAN JOURNAL OF ZOOLOGY-REVUE CANADIENNE DE ZOOLOGIE* 77.8, pp. 1175–1184. ISSN: 0008-4301. DOI: `10.1139/cjz-77-8-1175`. URL: `http://wm.summon.`

serialssolutions.com/2.0.0/link/0/eLvHCXMwnV1NS8QwEB1WRRDEb7GuQg_
irdo2bdLiQWRx8eZFL15KmqSwwn64rSz6651J27WKiHhum5RMMvOSvHkDwMIL3_
vmE6LAL5goYtKnCiT3i1Cj7Q1nMi1kUZivqVbQqkIO1m6dpPXceqro0PySMkQRDghxPXvxqIoU3bY2JTVW
TMXAR1SoqfeBFr6sAsY9EnwPwhBNlwM9yua6qWVqWQWCaLcYd43VFw_O-P78BWAOTdm3rm7ELPTPZg_
Wlqj9n34eoe5zi2h5HNnUs9mlaGiObV_M3NR7JORxOXJL4RqnqlLaWD9nXLmpR4AI_
D24fBndcUWvAUBvzASwi0SMH9lJFqMdOI4bhQEvdmeZTmvEh1qBKjQh3qVMfGRBJxWxrxWGphZMEOYVMSI
VKITjhWopEYlw2Dpy3Y5_NamGNzG5IWJqp5_dMiCwhJfLYgX47jJnM6fxHVWVGG6kYfUv3abP6ymw5wg6c
mpOobcYfwC2kvED.

Robinson, WS (2009). "Ecological Correlations and the Behavior of Indi-
viduals". In: *International journal of epidemiology* 38.2, pp. 337–341.
ISSN: 0300-5771. DOI: 10.1093/ije/dyn357. URL: http://wm.summon.
serialssolutions.com/2.0.0/link/0/eLvHCXMwjV1LS8NAEB6KBxGKj_
qKVVmw19gku9kkRwktehBBFLyFfYJCo7SV4r93t5vURLR6zmZZJdjYz32Tm-xYAR5eB_
8On0ECnLGYck9gGTK1UmGiquA4jJqmQbaoVDH4p6Gd4-PyihvKjxPGSQU6wTbju755WlQNCHOTFQeDHSRL
DLH0PROquOdyCvyTuu22QxaTRgN5Qc177ALmxXqBNduW2yBx1V9mDztqqr96Dr_
t4hR0rah2QkaqeIcnt-R9Uxh1gpkYGMqJJVnKJXjW5WnK7ZATyORw_5tV8dseALK9HpYy0YVSKW3MBGbfB
hC6zrfjlfEnZk8eAFMachQnmyoA1wk0KzKnmkU4kVykTmQcX9eIXb05So3ClcFyYxSnc4nhwbuyydkC_
NllRfXezIrLpp_GboQdHzopfM2TWFxHqwaBp1tV1GxsoSTO7c4yr8yD8z7C8UlK3CgLzk7-
euQ9bdTUqCE9hYz59V2fQWUw-AfbK7-M.

SELVIN, HANAN C. (1958). "Durkheim's Suicide and Problems of Empiri-
cal Research". In: *American Journal of Sociology* 63.5, p. 607. ISSN: 0002-
9602. URL: http://wm.summon.serialssolutions.com/2.0.0/link/
0/eLvHCXMw3V1dS8MwFL1MnwQRP9E5ISDMp0rbrE334EOZlQkyRKevo2kSLNJN1g7x39s0TVoV-
gN8baEpOXDuvck95wJg99q2fnGCb4sg9mKKR54MmIJzhwifU-G4MfMT9lNqZayfm2f_
Afjbzfr9jcshySQveSFNUmY0AXJ2TNW7EWUfqTIH0a137SzVX0OO0lWjbDFnMtHD633VJDANZ-

GsPm5lSlHnBU27XsuMtqsBRBNnWewo4uSaK4nlYTXtU5NpzVZp-7q6YkbfJkPpZp6xNClu-
NJ6ed4qS-VAFsxPk8eW5Zft6pJFLvgnRlZxbL4Pe_UWoFDt9AHO-PIQ-mY7OBApiTNSjitfR-
BpCK5yVAOASgCQBgCtBDIAIA3AMVzeRfPJ1NI_sqjhzhdldB9LCOGP4BPYjaUoYVlU4kV2CohjTGOHYMrL
Ab3P7BvEYB4u.

Weng, Qihao (2014). *Global Urban Monitoring and Assessment through Earth Observation.* en. CRC Press. ISBN: 978-1-4665-6450-3.

Willmott, C.J. and K Matsuura (2001). *Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1950-1999).* URL: `http://climate.geog.udel.edu/~climate/html_pages/README.ghcn_ts2.html`.

Zhu, Jun et al. (2004). "Combined mapping of soil properties using a multi-scale tree-structured spatial model". In: *Geoderma* 118.3, pp. 321–334. ISSN: 0016-7061. DOI: 10.1016/S0016-7061(03)00217-9. URL: `http://wm.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwpV1LT8MwDLYGu4AQb8R4SDlwgEOhjy
HzvNxkAgEFx6iJqktS3bST5_AcjS8zj65BMoarqsk2iVWlu1FSplM4OG0hOZFMavGxdKrerSmACyDJGg9v
icwpOJHRcpqtlqDJ5kuG3ry8ub_tzY8W8jhwNSYy4g7vZT31IL7xNM7O_DiR-
i5gfRGnfEzqb4CeVfbUUJTX8QI6e4Hm8R9_twnrIWEV3fq9LWjg4zasdofPgbQDd2BAToUW2GjFWDPdw1A
1PzNnq2CA_VBo4RGMOZRMAwWfiEc1gy0NZMWU4d0Omb-eZxcG_au73nUUrmuIKkobswhV7BKXt6WrnHTSx
OLL_-w-CM3sQGidS51tGOlTkPoq107jqkql0y04nymnnNT0HOU7bo2fLLUyzkovt1K1oJipsPwg-
5JCyE9dTxZVPp-PO2RapyhZ-IS6BclvXusFCnamHng5-PtXHcJKDSbiXaEjWCad4TE0XsdvQPELRw.`