AIDDATA
A Research Lab at William & Mary

WORKING PAPER 38

April 2017

geoSIMEX: A Generalized Approach To Modeling Spatial Imprecision

Daniel Runfola
Institute for the Theory and Practice of International Relations, AidData, William and Mary

Robert Marty
Institute for the Theory and Practice of International Relations, AidData, William and Mary

Seth Goodman
Institute for the Theory and Practice of International Relations, AidData, William and Mary

Michael Lefew
Institute for the Theory and Practice of International Relations, AidData, William and Mary

Ariel BenYishay
Institute for the Theory and Practice of International Relations, AidData, William and Mary

Abstract

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 while adjusting for spatial imprecision. We illustrate this approach through a case study leveraging a novel, publically available dataset recording the location of Chinese aid in Southeast Asia at varying levels of precision. Using a difference-in-difference modeling approach, we integrate Chinese aid information with satellite derived data on vegetation (NDVI) to examine if Chinese aid has caused an increase or decrease in vegetation. Following multiple approaches which do not incorporate spatial imprecision, we find 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 is found.

Author Information

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

Robert Marty
Institute for the Theory and Practice of International Relations, AidData, William and Mary

Seth Goodman
Institute for the Theory and Practice of International Relations, AidData, William and Mary

Michael LeFew
Institute for the Theory and Practice of International Relations, AidData, William and Mary

Ariel BenYishay
Department of Economics, William and Mary

Acknowledgments

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, Virginia Port Authority, Virginia Commonwealth Technology Research Fund, and the Office of Naval Research. The authors would also like to thank Ben Dykstra and Miranda Lv for their valuable insights.

Software

Accompanying this analysis, we provide a new software suite for researchers interested in (a) retrieving and integrating spatially-explicit data sources with varying levels of precision, and (b) modeling relationships using this data in a way which corrects for spatial imprecision in the R environment. This suite - the *geo* framework - is made available online, along with all documentation and code, at http://geo.aiddata.org. The case study presented in this paper leverages two components of the geo framework - the geo(query) tool - http://geo.aiddata.org/query - and the geoSIMEX R package - https://github.com/itpir/geoSIMEX. The *geo(query)* tool leverages a high performance cluster computing environment to enable dynamic requests and downloads of spatial data by individuals with limited background in GIS, outputting file formats (i.e., CSV) compatible with a wide range of statistical programs. The geoSIMEX R package is designed explicitly for a large body of researchers examining the impacts of international aid on environmental outcomes, but the main functions are generally adaptable to other use cases. Code for both tools is provided and made available under an open source license; extensive documentation on the use of this software, as well as the ability to download datasets customized for use with this software, can be found at http://geo.aiddata.org.

Contents

1. Introduction 1

1.1 Literature 1

2. Methods and Data 2

2.1 Methods 2

2.1.1 geoSIMEX 2

2.1.2 Simulations 5

2.1.3 Chinese Aid in Southeast Asia 6

2.2 Data 6

3. Results 7

3.1 Simulations 7

3.2 Chinese Aid in Southeast Asia 8

4. Discussion 8

4.1 geoSimex 9

5. Conclusion 10

6. Tables 11

7. Figures 15

References 18

1 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., Bare, Kauffman, and Miller 2015; Buntaine, Hamilton, and Millones 2015; Gallo and Goodchild 2012; Andam et al. 2008; Buchanan et al. 2016; BenYishay et al. 2016; Runfola and Napier 2016; Runfola et al. 2015; Landuyt et al. 2015; Chen et al. 2011; Ligmann-Zielinska and Jankowski 2014; Saint-Geours et al. 2014; Schluter and Ruger 2007). 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 Southeast Asia. We use this case study to illustrate the importance of including information on spatial imprecision into analyses. Finally, we introduce the *geo(query)* tool - with which all data used in this analysis can be retrieved - as well as an accompanying R package - geoSIMEX - for users seeking to incorporate information on spatial imprecision into analyses.

1.1 Literature

Past literature has shown that uncertainty in the locations of where measurements are taken can produce biased estimates in empirical analyses (Perez-Heydrich et al. 2013; Rettie and McLoughlin 1999). For example, Perez-Heydrich et al. 2013 show that regression coefficients can 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 radius of the measured location. One frequently cited "best practice" to overcome this challenge 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 - or, less - spatial uncertainty (Runfola et al. 2014; Giner et al. 2014; Perez-Heydrich et al. 2013). Yet another is to only use information for which exact (or, otherwise very precise) geographic information is known (Runfola and Napier 2016; Dreher et al. 2015; Runfola and Hughes 2014).

There are many limitations, assumptions, and biases that these approaches incur. Most predominant are the challenges highlighted by the large and well-established body of literature illustrating that analyzing data at different levels of aggregation can produce different regression coefficients and correlation coefficients; given the established nature of this literature we do not provide a full review here, but suggest a number of resources for readers new to employing spatial information (see Clark and Avery 1976; Goodchild 2001; Selvin 1958; Gotway and Young 2002; Gehlke and Biehl 1934; Cramer 1964; Pogson and Smith 2015; Gupta and Tarboton 2016). As a simple example of this concern, using data from the 1930 US census, Robinson 2009 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. A number of techniques have been proposed to address such biases, but most rely on additional

assumptions (or, covariates) to aid in an effective disaggregation of data to finer scales (i.e., Gotway and Young 2002; Zhu et al. 2004; O'Loughlin 2000; Wong 2004).

The simulation and extrapolation method (SIMEX) provides a solution to address measurement error in covariates (Wang et al. 1998; Kuchenhoff, Mwalili, and Lesaffre 2006; Li and Lin 2003; Cook and Stefanski 1994) with a minimal set of assumptions and no additional covariate information, but has not previously been applied to spatial imprecision. In traditional application, 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:

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

where X is the measured covariate, U is the variance of the measurement error and λ 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 λ (e.g., 0.5, 1, 1.5, and 2) to estimate models with simulated amounts of additional measurement error. Next, SIMEX estimates a trend between λ and the coefficient on X, and uses the trend to extrapolate back to $\lambda = -1$ (point of no measurement error). In the next section, we describe a novel deriviation of this approach - geoSIMEX - which adapts SIMEX to the case of geographic imprecision.

2 Methods and Data

2.1 Methods

Two different approaches are followed to illustrate the validity and applicability of the geoSIMEX process. First, we use a monte carlo simulation procedure to examine the relative accuracy of geoSIMEX as contrasted to other procedures. Second, we apply geoSIMEX to a case study of the impact of Chinese aid on environmental outcomes in Southeast Asia to provide an illustrative example of when the geoSIMEX approach might lead a researcher to a different conclusion than traditional approaches.

2.1.1 geoSIMEX

In this section, we detail the geoSIMEX approach using an illustrative example, in which we solve the following simplified equation, with hypothetical districts as units of analysis:

$$NDVI = \theta * \text{Chinese aid} + \epsilon \tag{2}$$

in which Chinese aid is measured with spatial imprecision (due to, for example, limited documentation on which district aid is being sent to). Through using geoSIMEX we account for this spatial imprecision to accurately estimate the model coefficient θ, as well as relevant metrics of significance.

In figure 1, we present a hypothetical country with sixteen districts for which we seek to solve equation 2. 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 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 longitude). Due to weaker project documentation, location B has a precision level indicating that it could have been allocated anywhere in the region that includes districts 5, 6, 7, and 8, but the exact location is unknown. Project C has very uncertain spatial information, such that it may be anywhere in the country.

The area (in square kilometers) in which each project location may have been allocated is summarized in table 2. Using the spatial overlap between each unit of analysis (the sixteen districts) and the known area an aid project might exist in, we calculate a probability that each district contains a given project[1]:

$$V_t = \sum_{S}^{s=1} U_s \left(\frac{a_{st}}{\sum_{T}^{t=1} a_{st}} \right) \tag{3}$$

These probabilities are used in the below steps of the geoSIMEX procedure to estimate model parameters while accounting for spatial imprecision.

Before geoSIMEX is run, users must calculate the initial level of spatial imprecision in a given dataset, defined by λ. To reflect imprecision across a given set of aid projects and a set of units of analysis (i.e., districts), we calculate (λ) following:

$$\lambda = \frac{\sum_{i}^{P} Area\ of\ Coverage_i}{\sum_{i}^{P} Total\ Possible\ Area\ of\ Coverage_i} \tag{4}$$

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, λ would resolve to 0—indicating zero spatial imprecision. If spatial data was only available for the entire study area (e.g., aid provided to a country without any indication of where the project was allocated), λ would resolve to 1—indicating 100% spatial uncertainty. In practice, combinations of different levels of precision in the documentation of individual aid projects result in λ values between these two extremes, resulting in λ between 0 and 1, in which larger values indicate higher spatial imprecision across all measurements.

After λ is known, the first step of geoSIMEX involves estimating a naive model (which can be of variable functional forms; for illustration we use ordinary least squares regression) using the source data. In this

[1] Here, 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).

example case, for each unit of observation (i.e., district), the total dollars of Chinese Aid in equation 2 is estimated by equally spreading aid according to geographic overlap - i.e., dollars from each project are weighted by the size of each district a given dollar could fall in to, and the final value is the weighted sum across all projects for each district. In figure 2, we provide an example of the geoSIMEX procedure applied to a dataset with an initial λ value of 0.4; figure 2A illustrates the result from this step for the example dataset. In this figure, the x-axis represents the λ value (spatial imprecision) for a dataset, with higher values indicating more imprecision. The y-axis represents the estimated value of θ in equation 2. In 2a, the orange line represents the 95% confidence interval of the coefficient on aid in the naive model fit in step 2, this estimate is plotted at $\lambda = 0.4$, following the estimate of spatial imprecision calculated using equation 4. The horizontal black line represents the true model coefficient generated for this example (θ = 1), which the naive model fails to capture.

In the second step, additional imprecision is simulated by randomly decreasing the spatial precision of information (in this case, the spatial precision with which Chinese aid projects are known), and re-calculating the total dollars of aid for each unit of observation according to the updated areas-of-overlap between each unit and aid project. For example, a project that has a measurement with an exact latitude and longitude will randomly be assigned a lower level of spatial 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 θ parameter, standard errors of the model, and λ value for a given permutation are saved. In figure 2b, the black points represent individual iterations, with the saved model coefficients (θ, y axis) and their associated λ (x axis) values.

The third step subdivides this set of iterations into four equally-sized bins based on the level of spatial uncertainty (λ) of the aid variable (e.g., if λ values range from 0.4 to 1, coefficients are separated into bins of 0.40-0.55, 0.55-0.70, 0.70-0.85 and 0.85-1.00). Average coefficient and λ values are calculated within each bin, represented as red dots in figure 2c. The fourth step (2d) fits a quadratic trend to the average coefficient and lambda values estimated in step 3. The trend is then extrapolated back to $\lambda = 0$, providing an estimate of θ 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 using a bootstrap procedure. An iterative procedure is followed in which a single point from each bin (defined in step 3) is sampled, a quadratic trend is fit on the resulting values, and the trend is extrapolated back to λ = 0. This process is repeated iteratively to capture as many permutations as is computationally feasible (with the number of iterations defined as R). In figure 2e, each blue line represents one extrapolated trend and λ = 0 estimate. Following the approach outlined in Burnham and Anderson 2002, we use this information to explicitly quantify both the original standard errors and the additional error from spatial imprecision:

$$var(\hat{\bar{\theta}}) = \sum_{i}^{R} \frac{1}{R}\{var(\hat{\theta}_i) + (\hat{\theta}_i - \hat{\bar{\theta}}_i)^2\} \tag{5}$$

where R is the number of extrapolated coefficients (in this example, 1000). $var(\hat{\theta}_i)$ 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. $(\hat{\theta}_i - \hat{\bar{\theta}}_i)^2$ captures the remaining variance attributable to spatial uncertainty, based on the range of possible θ outcmes found in the extrapolation procedure (step 4).

2.1.2 Simulations

We employ a monte carlo simulation procedure to examine the accuracy of geoSIMEX compared to other approaches to handle spatial imprecision. Each simulation follows 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) a spatial area, and (b) 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. A simulated measurement of NDVI is generating according to equation 2 (defined as the number of aid projects plus random error).

4. Each aid project is given a code indicating the spatial precision that a 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. For each iteration, based on the spatial precision our hypothetical researcher would have access to 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 receiving aid.[2]

6. Equation 2 is estimated using 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 (dropping out other information).

7. Equation 2 is estimated using models that seek to incorporate spatial imprecision: geoSIMEX and a model averaging approach. In the model averaging approach, subcounties are assigned a probability of receiving aid according to their spatial size, and the expected value of aid is calculated iteratively according to that probability. 500 models are estimated, where the average coefficient is used with the standard error estimated using equation 5. This approach is contrasted to the geoSIMEX approach.

[2]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; this assumption could be modified according to dasymetric mapping approaches if desired by the researcher.

This process is repeated approximately 6000 times, and the rate at which each modeling approach is able to capture the true relationship between aid and NDVI is contrasted along two dimensions. First, we examine the percent of times that each approach captured the true relationship between Chinese aid and NDVI (θ) within a 95% confidence interval. Second, we assess the percent of iterations a hypothetical researcher would have identified these results as being statistically significant at the 95% level.

2.1.3 Chinese Aid in Southeast Asia

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. This study leverages a novel dataset on the location of Chinese international aid in Southeast Asia 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); all data sources are summarized in section 4.2. Employing geoSIMEX, we use this case study to illustrate the importance of incorporating information on spatial imprecision into analyses.

A difference-in-difference modeling strategy is followed, in which the average of NDVI before aid was allocated (pre-2004) is contrasted to the average of NDVI after aid was allocated (post-2011) for each of 351 districts (see section 4.2 for more information):

$$Y_i = \beta_0 + \theta * Aid_i + \sum_{k=1}^{n} (\beta_k * X_i) \tag{6}$$

Where Y_i is the difference in the average forest loss post-2011 and pre-2004 in district i; Aid_i is the number of aid projects in each district; X_i is a vector of control variables; β_k is a vector of parameters for each beta covariate, and θ is the estimated impact of aid. We calculate the initial λ value for our dataset following equation 4, and simulate additional spatial imprecision by allocating aid to increasingly coarse units of administration. We repeat this analysis using three alternative modeling approaches to contrast their outcomes to geoSIMEX: (1) an ordinary least squares model in which imprecision is ignored (representative of a traditional case of the ecological fallacy), (2) a linear model in which all coarse-resolution data is removed from the dataset, and (3) a Monte Carlo model averaging approach in which hundreds of equally-probable worlds are generated based on the initial dataset, and the average beta coefficient and standard errors are reported.[3]

2.2 Data

In this analysis, we examine the impact of Chinese aid in Southeast Asia, explicitly focusing on second-level administrative units within Cambodia, Laos, Myanmar, Thailand and Vietman. To conduct this analysis, we leverage a dataset on the location of Chinese aid derived through a methodology designed to Track Underreported Financial Flows (TUFF; Strange et al. 2015). Covariate data is collected from a variety

[3]It should be noted there are many approaches to model averaging, including AIC weighting, that are not tested in this paper. It is possible some of these alternative approaches may outperform the Monte Carlo procedure presented here.

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. (Willmott and Matsuura 2001)

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). (Farr et al. 2007)

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.

3 Results

3.1 Simulations

Table 3 shows results from the simulation analysis. Results are subset according to the spatial imprecesion of the simulated aid data, with each row of results representing simulations grouped according to the simulated uncertainty (the final row summarizes across all results). The naive OLS model using the expected value of aid (i.e., ignoring spatial imprecision) captures the true coefficient about 50% of the time within a 95% confidence interval when there is low spatial uncertainty ($\lambda < 0.3$), and the ability of OLS to capture the true coefficient declines as spatial imprecesion increases. The naive model that omits aid projects measured with spatial uncertainty captures the true coefficient 60% to 70% of the time at a

95% confidence interval, 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 imprecesion increases the ability for the model averaging approach to capture the true coefficient decreases. Under low levels of spatial imprecesion 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 imprecesion. At low levels of spatial imprecesion the geoSIMEX model captures the true coefficient 90% of the time. 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 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 result in geoSIMEX capturing the true relation and statistical significance 0% of the time.

3.2 Chinese Aid in Southeast Asia

We examine the model results from our analysis of the impact of Chinese Aid in Southeast Asia to provide an illustrative example of how accounting for spatial imprecision can lead to substantively different conclusions than alternative modeling approaches. Table 4 summarizes these findings. In (1) a linear model that ignores spatial imprecision (i.e., commits an ecological fallacy), (2) a linear model that omits imprecise spatial information, and (3) a Monte Carlo model averaging approach, findings suggest that Chinese Aid has contributed to deforestation. Specifically, an additional aid project within a district is associated with a 1.1 to 1.7 percentage point decrease in NDVI ($p < 0.05$). However, when the spatial imprecision of the data is explicitly modeled (4), we find that insufficient evidence exists to determine if Chinese Aid has had a positive or negative impact on vegetation ($p > 0.1$).

Further, leveraging geoSIMEX we seperately estimate the contribution of (1) spatial imprecision and (2) unexplained model variance to the standard errors estimated for θ. Here, we find that spatial imprecision contributes 98.5% of the variance around θ, while unexplained model variance contributes 1.5%.

4 Discussion

These results highlight the importance of incorporating spatial imprecision into analytic approaches which seek to establish causal relationships using spatial data. Under current "best practice" approaches, a researcher would find that Chinese aid has had a statistically significant, detrimental and causal impact on deforestation in Southeast Asia, even after controlling for a wide variety of potential confounding variables. This result - which could have strong policy ramifications - is illustrated to have insufficient evidence to support it when the geoSIMEX procedure is followed, largely due to the relatively imprecise nature of the measurements of aid. Failing to account for spatial imprecision can thus bias results in key ways, and lead researchers to conclusions that may be inaccurate. Further, we find nearly all (98.5%) of the unex-

plained variance can be ascribed to spatial imprecision; this offers compelling evidence that the true relationship between Chinese aid and deforestation could be better understood if more precise spatial information was collected.

4.1 geoSimex

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.

Our results suggest that geoSIMEX can provide a solution to cases in which researchers have known spatial imprecision in available data. As the precision of source data decreases, the ability of geoSIMEX to accurately estimate θ within a 95% confidence interval was significantly better than the alternative modeling approaches we examined. This is better in line with the expectations of researchers employing linear models - i.e., under traditional assumptions researchers expect that - at the 95% confidence interval - the true relationship will fall within that interval 95% of the time. Even under cases of exceptionally good spatial information - λ less than 0.3 - this was only true for 53% of the linear models, and 63% of the models that leveraged exact spatial information. Improving on these results, the geoSIMEX procedure resulted in approximately 87% of cases falling within the expected confidence interval. Increasing the number of simulations beyond the 6000 used here could indicate further improvement in the geoSIMEX procedure; ongoing work is identifying opportunities to optimize this computational challenge.

Despite the relative improvement of geoSIMEX in capturing the true model relationship, this comes at a cost of increased bands of uncertainty. While we argue this is more reflective of the input data - i.e., higher uncertainty should be expected when the spatial precision of input data is low - it practically results in a lower likelihood of detecting statistical significance. This is reflected by the rapid drop-off in the geoSIMEX procedure's rate of identifying both the true coeficient and significance in table 3.

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 λ approaches 0 (perfect precision relative to the units of observation), geoSIMEX will provide increasingly accurate estimates as it has an increased number of λ units to derive data for - i.e., there are more variable simulations across observations of λ. This limitation 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, 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 - i.e., datasets in which imprecision is not ascribed to known units such as administrative zones - the application of geoSIMEX is currently not feasible due to a lack of a relevant information regarding how data should be simulated at higher levels of λ. 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.

5 Conclusion

In this piece, we introduce a flexible method to account for spatial imprecison in datasets for the purpose of fitting models in the case of spatial imprecision in measurements - geoSIMEX. We illustrate 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, though at the tradeoff of a lower ability to identify statistical significance in relationships. 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 statistically significant, detrimental impact on forest cover. By accounting for spatial imprecision, we illustrated that there was insufficient evidence to draw this conclusion. We use these findings to argue for the importance of incorporating spatial imprecision into analyses. Finally, we introduce the geo(query) software - with which all data used in this analysis can be retrieved - as well as an accompanying R package - geoSIMEX - for users seeking to incorporate information on spatial imprecision into analyses.

6 Tables

Table 1: Data sources used in this analysis.

Data Sources	
Data Name	**Source**
Chinese Aid Locations	AidData[4]
Gridded Population of the World	Center for International Earth Science Information Network[5]
Nighttime Lights	Defense Meteorological Satellite Program[6]
Precipitation and Temperature	University of Delaware (Willmott and Matsuura 2001)[7]
Urban Travel Time	European Commission Joint Research Centre[8]
Distance to Rivers	World Wildlife Fund [9]
Vegetation	NASA LTDR[10]
Distance to Roads	CIESIN gRoads[11]

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

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

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

[7] 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.

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

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

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

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

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

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 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. "Sig." refers to significant at the 95% level.

Table 4: Impact of Infrastructure Aid on Forest Loss

	Forest Loss			
	Naive	Naive (Low PC)	Model Avg	geoSIMEX
	(1)	(2)	(3)	(4)
Aid	1.745***	1.126**	1.694***	2.218
	(0.523)	(0.553)	(0.535)	(4.595)
Air Temp (Max)	0.107	0.116	0.107	0.099
	(0.248)	(0.248)	(0.248)	(0.264)
Air Temp (Min)	1.037**	0.997*	1.037**	1.014
	(0.521)	(0.522)	(0.522)	(0.648)
Air Temp (Avg)	3.487***	3.572***	3.500***	3.498***
	(0.925)	(0.926)	(0.925)	(1.247)
Precip (Max)	0.286***	0.287***	0.286***	0.287***
	(0.029)	(0.029)	(0.029)	(0.030)
Precip (Min)	0.007**	0.008***	0.008**	0.007
	(0.003)	(0.003)	(0.003)	(0.005)
Precip (Avg)	−0.032**	−0.035**	−0.033**	−0.032
	(0.015)	(0.015)	(0.015)	(0.020)
lnyx	0.0001	0.0001	0.0001	0.0001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Nighttime Lights	−0.021	−0.021	−0.021	−0.018
	(0.033)	(0.033)	(0.033)	(0.034)
Constant	1.999***	2.035***	2.004***	1.989***
	(0.335)	(0.335)	(0.335)	(0.389)
Observations	1,869	1,869	1,869	1,869

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors in parentheses

7 Figures

Figure 1: Hypothetical example of spatial imprecision in aid allocation.

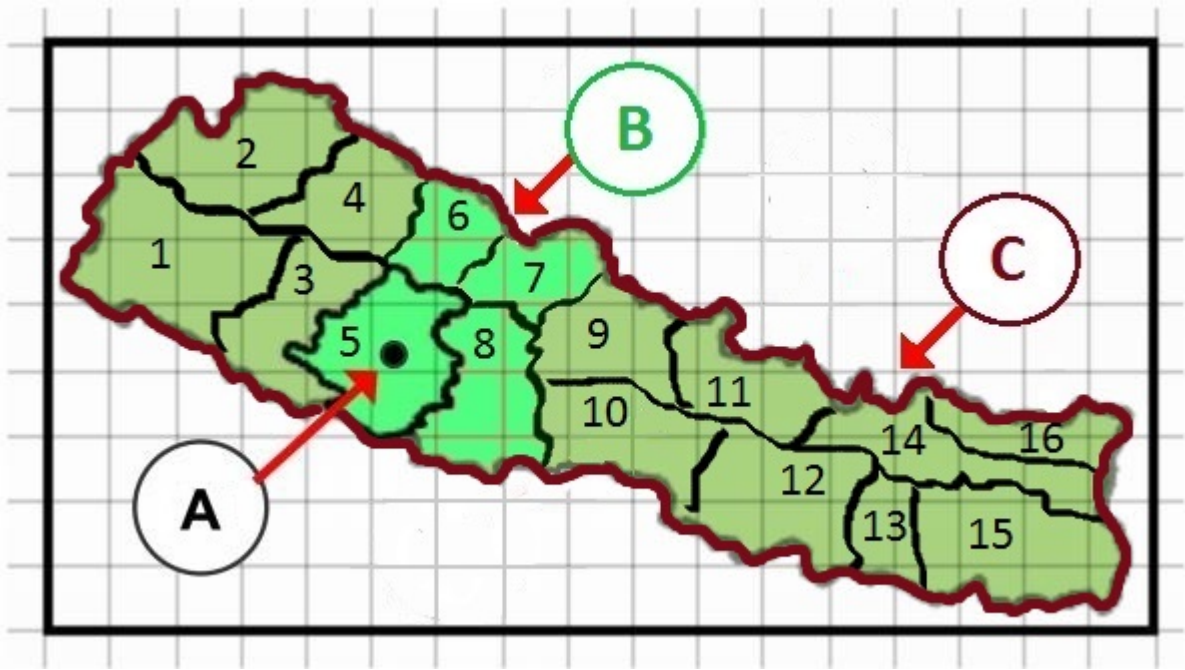


Figure 2: Steps of the geoSIMEX procedure.

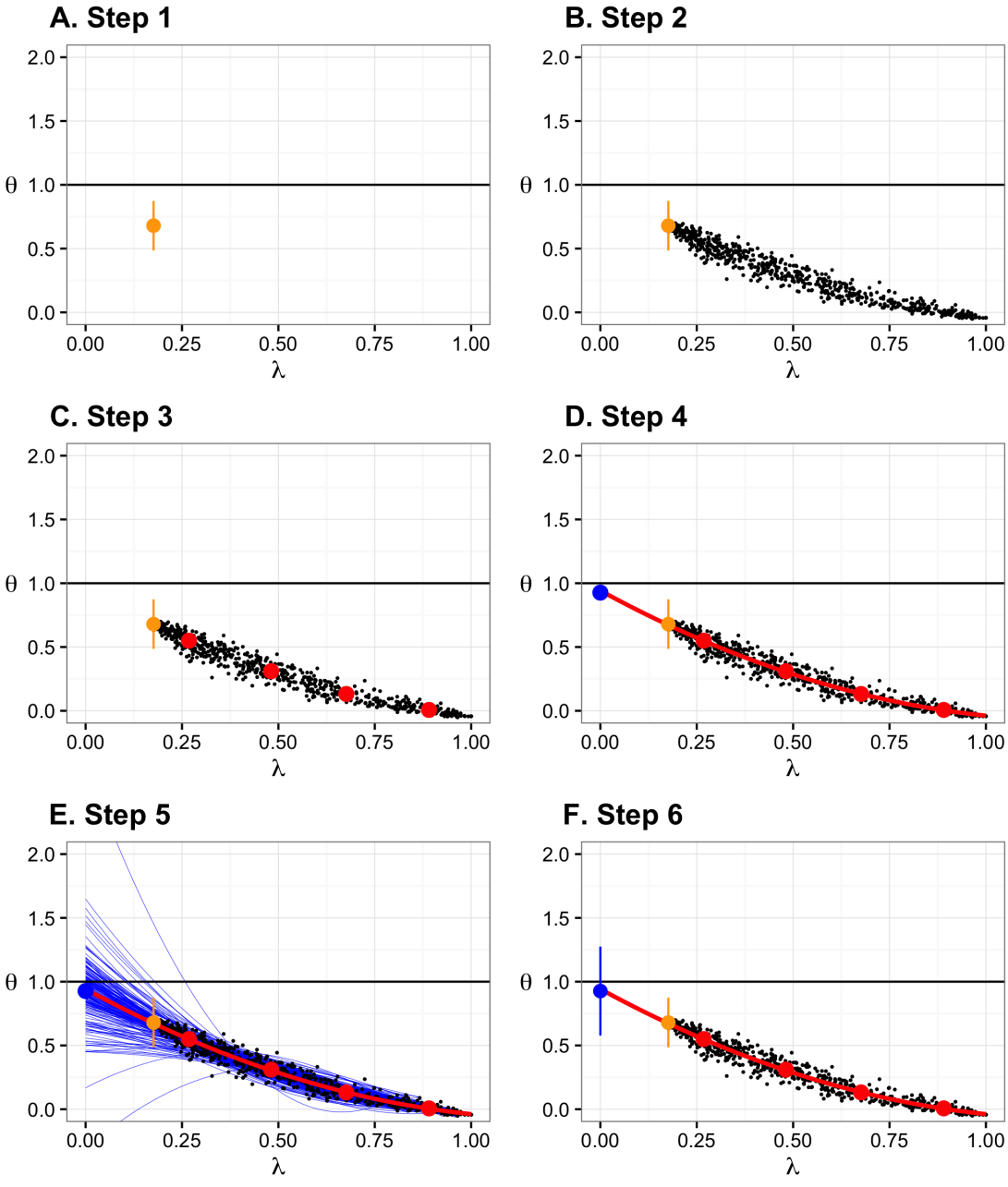
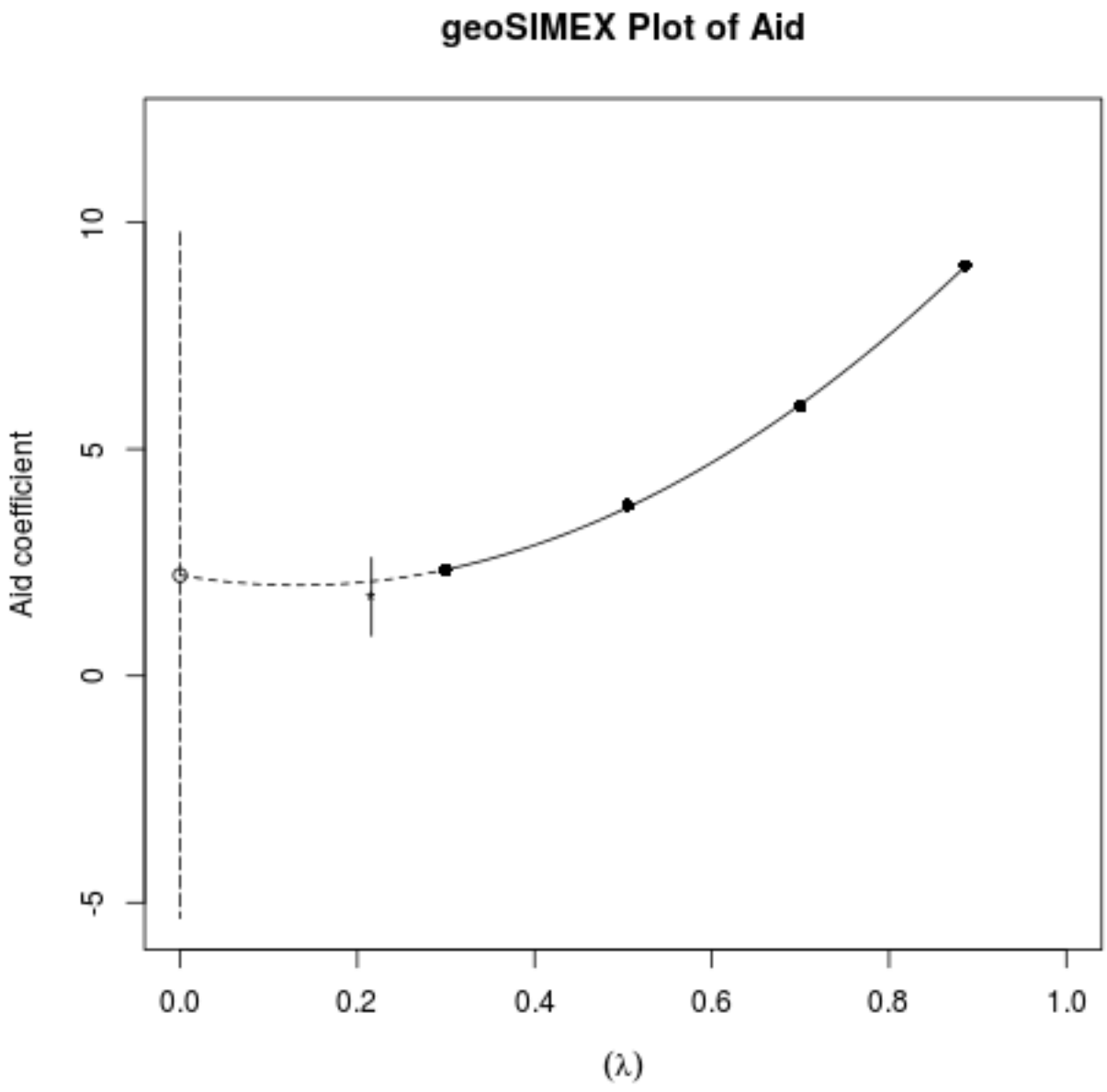


Figure 3: geoSIMEX Plot of Aid.



References

Andam, Kwaw S. et al. (2008). "Measuring the effectiveness of protected area networks in reducing deforestation". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.42, pp. 16089–16094. ISSN: 1091-6490. DOI: 10.1073/pnas.0800437105.

Bare, Matthew, Craig Kauffman, and Daniel C. Miller (2015). "Assessing the impact of international conservation aid on deforestation in sub-Saharan Africa". en. In: *Environmental Research Letters* 10.12, p. 125010. ISSN: 1748-9326. DOI: 10.1088/1748-9326/10/12/125010. URL: http://stacks.iop.org/1748-9326/10/i=12/a=125010 (visited on 05/23/2016).

BenYishay, Ariel et al. (2016). "Indigenous Land Rights and Deforestation: Evidence from the Brazilian Amazon". In: URL: http://aiddata.org/sites/default/files/wps22_indigenous_land_rights_and_deforestation.pdf (visited on 11/21/2016).

Buchanan, Graeme M. et al. (2016). "The Impacts of World Bank Development Projects on Sites of High Biodiversity Importance". In: URL: http://aiddata.org/sites/default/files/wps20_world_bank_biodiversity_0.pdf (visited on 11/21/2016).

Buntaine, Mark T., Stuart E. Hamilton, and Marco Millones (2015). "Titling community land to prevent deforestation: An evaluation of a best-case program in Morona-Santiago, Ecuador". In: *Global Environmental Change* 33, pp. 32–43. ISSN: 0959-3780. DOI: 10.1016/j.gloenvcha.2015.04.001. URL: http://www.sciencedirect.com/science/article/pii/S0959378015000503 (visited on 03/16/2016).

Burnham, KP and DR Anderson (2002). "Information and likelihood theory: a basis for model selection and inference". In: *Model selection and multimodel inference: a practical information-theoretic approach*, pp. 49–97.

Chen, H. et al. (2011). "Uncertainty analysis in a GIS- based multi- criteria analysis tool for river catchment management". In: *Environmental Modelling and Software* 26.4, pp. 395–405. DOI: 10.1016/j.envsoft.2010.09.005.

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.

Cook, J.R. and L.A. Stefanski (1994). "Simulation-extrapolation estimation in parametric measurement error models". In: *Journal of the American Statistical Association* 89.428, pp. 1314–1328. URL: http://www.jstor.org/stable/2290994?seq=1#page_scan_tab_contents.

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.

Dreher, Axel et al. (2015). *Aid on Demand: African Leaders and the Geography of China's Foreign Assistance*. SSRN Scholarly Paper ID 2630152. Rochester, NY: Social Science Research Network. URL: https://papers.ssrn.com/abstract=2630152 (visited on 10/14/2016).

Farr, Tom G. et al. (2007). "The Shuttle Radar Topography Mission". en. In: *Reviews of Geophysics* 45.2. ISSN: 8755-1209. DOI: 10.1029/2005RG000183. URL: http://doi.wiley.com/10.1029/2005RG000183 (visited on 11/21/2016).

Gallo, John and Michael Goodchild (2012). "Mapping Uncertainty in Conservation Assessment as a Means Toward Improved Conservation Planning and Implementation". In: *Society & Natural Resources* 25.1, pp. 22–36. ISSN: 0894-1920. DOI: 10.1080/08941920.2011.578119. URL: http://dx.doi.org/10.1080/08941920.2011.578119 (visited on 07/24/2014).

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.

Giner, Nicholas M. et al. (2014). "Creating Spatially-Explicit Lawn Maps Without Classifying Remotely-Sensed Imagery: The case of suburban Boston, Massachusetts, USA". In: *Cities and the Environment (CATE)* 7.1, p. 10. URL: http://digitalcommons.lmu.edu/cate/vol7/iss1/10/ (visited on 11/21/2016).

Goodchild, Michael (2001). "Models of Scale and Scales of Modelling". In: *Modelling scale in geographical information*. (Visited on 11/21/2016).

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.

Gupta, Avirup Sen and David G. Tarboton (2016). "A tool for downscaling weather data from large- grid reanalysis products to finer spatial scales for distributed hydrological applications". In: *Environmental Modelling and Software* 84, pp. 50–69. DOI: 10.1016/j.envsoft.2016.06.014.

Kuchenhoff, Helmut, Samuel M. Mwalili, and Emmanuel Lesaffre (2006). "A General Method for Dealing with Misclassification in Regression: The Misclassification SIMEX". en. In: *Biometrics* 62.1, pp. 85–96. ISSN: 1541-0420. DOI: 10.1111/j.1541-0420.2005.00396.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0420.2005.00396.x/abstract (visited on 10/14/2016).

Landuyt, Dries et al. (2015). "A GIS plug-in for Bayesian belief networks: Towards a transparent software framework to assess and visualise uncertainties in ecosystem service mapping". In: *Environmental Modelling and Software* 71, pp. 30–38. DOI: 10.1016/j.envsoft.2015.05.002.

Li, Yi and Xihong Lin (2003). "Functional Inference in Frailty Measurement Error Models for Clustered Survival Data Using the SIMEX Approach". In: *Journal of the American Statistical Association* 98.461,

pp. 191–203. ISSN: 0162-1459. DOI: 10.1198/016214503388619210. URL: http://dx.doi.org/10.1198/016214503388619210 (visited on 10/14/2016).

Ligmann-Zielinska, Arika and Piotr Jankowski (2014). "Spatially- explicit integrated uncertainty and sensitivity analysis of criteria weights in multicriteria land suitability evaluation". In: *Environmental Modelling and Software*. DOI: 10.1016/j.envsoft.2014.03.007.

O'Loughlin, John (2000). "Can King's Ecological Inference Method Answer a Social Scientific Puzzle: Who Voted for the Nazi Party in Weimar Germany?" In: *Annals of the Association of American Geographers* 90.3, pp. 592–601. ISSN: 0004-5608. DOI: 10.1111/0004-5608.00213. URL: http://dx.doi.org/10.1111/0004-5608.00213 (visited on 10/14/2016).

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.

Pogson, Mark and Pete Smith (2015). "Effect of spatial data resolution on uncertainty". In: *Environmental Modelling and Software* 63, pp. 87–96. DOI: 10.1016/j.envsoft.2014.09.021.

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.

Robinson, WS (2009). "Ecological Correlations and the Behavior of Individuals". In: *International journal of epidemiology* 38.2, pp. 337–341. ISSN: 0300-5771. DOI: 10.1093/ije/dyn357.

Runfola, Daniel Miller and Sara Hughes (2014). "What makes green cities unique? examining the economic and political characteristics of the grey-to-green continuum". In: *Land* 3.1, pp. 131–147. URL: http://www.mdpi.com/2073-445X/3/1/131/htm (visited on 11/21/2016).

Runfola, Daniel Miller and Ashley Napier (2016). "Migration, climate, and international aid: examining evidence of satellite, aid, and micro-census data". In: *Migration and Development* 5.2, pp. 275–292. ISSN: 2163-2324. DOI: 10.1080/21632324.2015.1022969. URL: http://dx.doi.org/10.1080/21632324.2015.1022969 (visited on 10/14/2016).

Runfola, Daniel Miller et al. (2015). "A multi-criteria geographic information systems approach for the measurement of vulnerability to climate change". In: *Mitigation and Adaptation Strategies for Global Change*, pp. 1–20. URL: http://link.springer.com/article/10.1007/s11027-015-9674-8 (visited on 11/21/2016).

Runfola, Daniel SM et al. (2014). "Using Fine Resolution Orthoimagery and Spatial Interpolation to Rapidly Map Turf Grass in Suburban Massachusetts". In: *International Journal of Geospatial and Environmental Research* 1.1, p. 4. URL: http://dc.uwm.edu/ijger/vol1/iss1/4/ (visited on 11/21/2016).

Saint-Geours, Nathalie et al. (2014). "Multi- scale spatial sensitivity analysis of a model for economic appraisal of flood risk management policies". In: *Environmental Modelling and Software*. DOI: 10.1016/j.envsoft.2014.06.012.

Schluter, Maja and Nadja Ruger (2007). "Application of a GIS- based simulation tool to illustrate implications of uncertainties for water management in the Amudarya river delta". In: *Environmental Modelling and Software* 22.2, pp. 158–166. DOI: 10.1016/j.envsoft.2005.09.006.

Selvin, Hanan C. (1958). "Durkheim's Suicide and Problems of Empirical Research". In: *American Journal of Sociology* 63.5, p. 607. ISSN: 0002-9602.

Strange, Austin M. et al. (2015). "Tracking Underreported Financial Flows China's Development Finance and the Aid–Conflict Nexus Revisited". en. In: *Journal of Conflict Resolution*, p. 0022002715604363. ISSN: 0022-0027, 1552-8766. DOI: 10.1177/0022002715604363. URL: http://jcr.sagepub.com/content/early/2015/09/18/0022002715604363 (visited on 10/18/2016).

Wang, Naisyin et al. (1998). "Bias Analysis and SIMEX Approach in Generalized Linear Mixed Measurement Error Models". In: *Journal of the American Statistical Association* 93.441, pp. 249–261. ISSN: 0162-1459. DOI: 10.1080/01621459.1998.10474106. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1998.10474106 (visited on 10/14/2016).

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.

Wong, David W. S. (2004). "The Modifiable Areal Unit Problem (MAUP)". en. In: *WorldMinds: Geographical Perspectives on 100 Problems*. Ed. by Donald G. Janelle, Barney Warf, and Kathy Hansen. Springer Netherlands, pp. 571–575. ISBN: 978-1-4020-1613-4 978-1-4020-2352-1. URL: http://link.springer.com/chapter/10.1007/978-1-4020-2352-1_93 (visited on 10/14/2016).

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.