The SLEUTH Wizard: Python scripts to automate the SLEUTH urban growth model

Alfonso Yáñez Morillo[1], Dr. Claire Jantz[1], Tiernan Erickson[2]

[1]Shippensburg University (ayanezmorillo@ship.edu; cajant@ship.edu), [2]US Census Bureau (tiernan.erickson@census.gov)

Abstract

SLEUTH (Clarke, Hoppen, and Gaydos 1997) is one of the more broadly applied models for the study of land use change and urban dynamics. When the model is being applied to a large region, it is often desirable to partition the study area into sub-regions, such as states, counties, or watersheds. This sub-regionalization greatly increases the workload, requiring the preparation of each sub-region's input data sets and run parameters, and then evaluating multiple output files for each sub-region. To solve this problem, we developed two Python scripts that automate much of the workflow, saving time and minimizing user error. To use the scrips, the user must first have all of the base data sets prepared for the entire study area. The first script uses arcpy to extract the information for user-specified sub-regions (e.g. counties or watersheds) and then stores it in the correct format, using the correct naming convention, in a directory system that is ready to use for SLEUTH. The second script, called SWizard, is able to perform calibration, validation, and prediction automatically depending on the user needs. To demonstrate the SWizard capabilities, we applied SLEUTH to the continental United States, using counties as our sub-regions, at a resolution of 360m per pixel. Extracting the input data for 3,109 counties took 1 hour 6 minutes, while running SWizard on a single Linux desktop computer took 19 hours and 23 minutes. We were thus able to model urban land change for the entire continental US in less than 24 hours.

Overview of SLEUTH & Impetus for this Study

SLEUTH is one of the more widespread cellular automata models to simulate land use/land cover change. It uses four growth rules (spontaneous, new centers, edge growth, and road-influenced growth) that are controlled by five parameters (diffusion, breed, spread, slope, and road) and requires a computationally and labor intensive calibration. Among its limitations, the model fails to match the growth pattern correctly when the study area is characterized by a heterogeneous and diverse urbanization pattern, since the same combination of parameters is applied for the entire area. To correct that issue, it is necessary to divide the area of interest in more homogeneous sub-regions (e.g. counties or watersheds) where urbanization pattern variability is minimized (Jantz et al. 2010).

In that context, carrying out SLEUTH simulations becomes a tedious and repetitive task highly prone to causing mistakes, with the subsequent waste of time and effort.

Figure 1

Figure 2a

```
Project Version 10
delaware
0.005
0      25     100
5      5      15
5      5      20
10     5      30
20     2      40
1      2      3      4

ulster;Philadelphia
0.04
0      25     100
5      5      15
5      5      20
10     5      30
20     2      40
10     20     30     40     50
END
```

Figure 2b

```
Conus test 5
0.005
END
```

Figure 1. *A* snapshot of part of the scenario file

Figure 2. (*a*) The structure of the settings file is organized to give all the required information to SLEUTH for a large number of regions. The first line is common for all the regions and indicates the location of the data, a text to identify the results, and (optionally) the number of Monte Carlo trials. The next 8 lines are repeated for all clusters of regions and represent, in order: Auxiliary Diffusion Multiplier, coefficients for calibration (Diffusion, Breed, Spread, Slope and Road), and the best fitted set of coefficients for prediction mode. The file finishes with an 'END' flag. (*b*) Settings file used for the calibration and prediction of the study area (see CONUS example below). Depending on the option chosen, not all lines are needed. Here, the script was run in automated mode for all counties, so neither coefficients nor counties' names were necessary.

SLEUTH Wizard (SWizard)

SWizard was developed to automate many of the workflow tasks involved with calibration, validation, and forecasting with SLEUTH. It is a script written in Python that runs in Linux, and intermediates between the user and SLEUTH. A text file, called *the settings file*, is used to tell the script what it has to do (fig 2). The power of this script is that it is useful for launching the SLEUTH program for a large number of scenarios in just one go, but is flexible so that the user can have complete control of the process or let the script do all the hard work.

```
(0) Aux_Diff_Multi coefficient calibration
(1) Only Scenario File for calibration
(2) Only Scenario File for prediction
(3) Calibration using Settings File
(4) Automatic Calibration
(5) Only Prediction
(6) Calibration and Prediction
(99) Exit
```

Figure 3. SWizard options

Automatic Calibration with SWizard

SWizard can replicate the commonly used "brute force" calibration technique to find a parameter combination with the best fit score. SLEUTH compares the prediction with the observational data to calculate a number of metrics to measure the model's goodness of fit. There is no consensus for choosing the best metric, and each metric measures different aspects of the model's performance. Jantz et al. (2010) and Jantz, Drzyzga and Maret (2014) focus on spatial pattern metrics and measures of the overall amount of development. In these examples, fractional difference metrics that compare simulated and observed pixels (urban land), edges, and clusters are selected and integrated in a new composite metric of goodness of fit.

Table 1. SLEUTH-3r provides a set of new fit metrics which compares the observed and modeled values. For each of them SLEUTH-3r calculates three versions: the algebraic difference of modeled and observed, the ratio, and fractional change, both referenced to the observed value. The metrics described below were used to calculate the new global fit metric. (Table extracted from Jantz et al. 2010)

Fit statistic	Definition
Pixels	Modeled urban pixels compared to actual urban pixels for each control year. Referred as "population" and as "area" in SLEUT's output files
Edges	Modeled urban edges pixels compared to actual urban edge pixels for each control year
Clusters	Modeled number of urban clusters compared to actual urban clusters for each control year. Urban clusters are areas of contiguous urban land using the 8-neighbor rule.

$$Global\ Fit\ Metric = \sqrt{area^2 + edge^2 + cluster^2}$$

During the calibration process, SWizard calculates the composite fit metric and selects one combination of parameters with the lowest value, which indicates that the simulation is a close match to the observed. Each selection is identified and the associated combination and metrics, including the composite fit metric, are stored in a csv file. Users have the ability to edit the Python script to formulate their own composite metric, such as the optimal SLEUTH metric (OSM) (Dietzel and Clarke 2007).

Testing the capabilities of SWizard: An Analysis of the Continental US

To test the capacity of the new script, we ran it for the entire continental United States (CONUS). For demonstration purposes, the county was selected as the spatial unit for analysis, resulting in a total of 3,109 counties to be assessed and modeled individually.

For this application, we relied on national-scale, public domain raster and vector data sets (table 2). Once maps for the whole area were made, the arcpy script was used to extract and store each county's data with the appropriate naming convention, to a directory structure ready to be read by SWizard (fig 4).

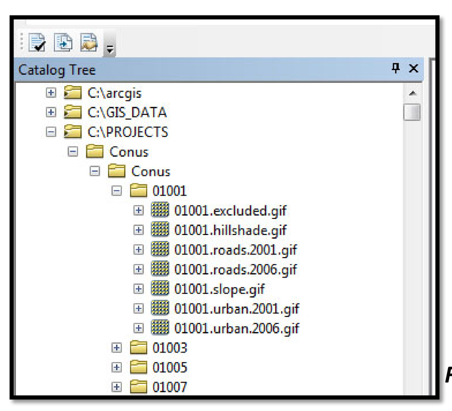


Figure 4. Directory structure accessible for Swizard.

Table 2. Input data sets; All data sets were rasterized at the same extent and grid system at a resolution of 360m.

Layer	Source	Description
Urban	National Land Cover Data	Urban (1): Impervious classes Not urban (0): rest of classes
Exclusions	USGS Protected Areas Database (2012), and National Wetlands Inventory (2014)	Completely excluded (100): water bodies and protected areas Partially excluded (80): Wetlands
Slope	The USGS 1/3 arc second digital elevation models	Percentage slope derived from DEM
Transportation	Census Bureau TIGER roads	S1100, S1200, S1400, Other

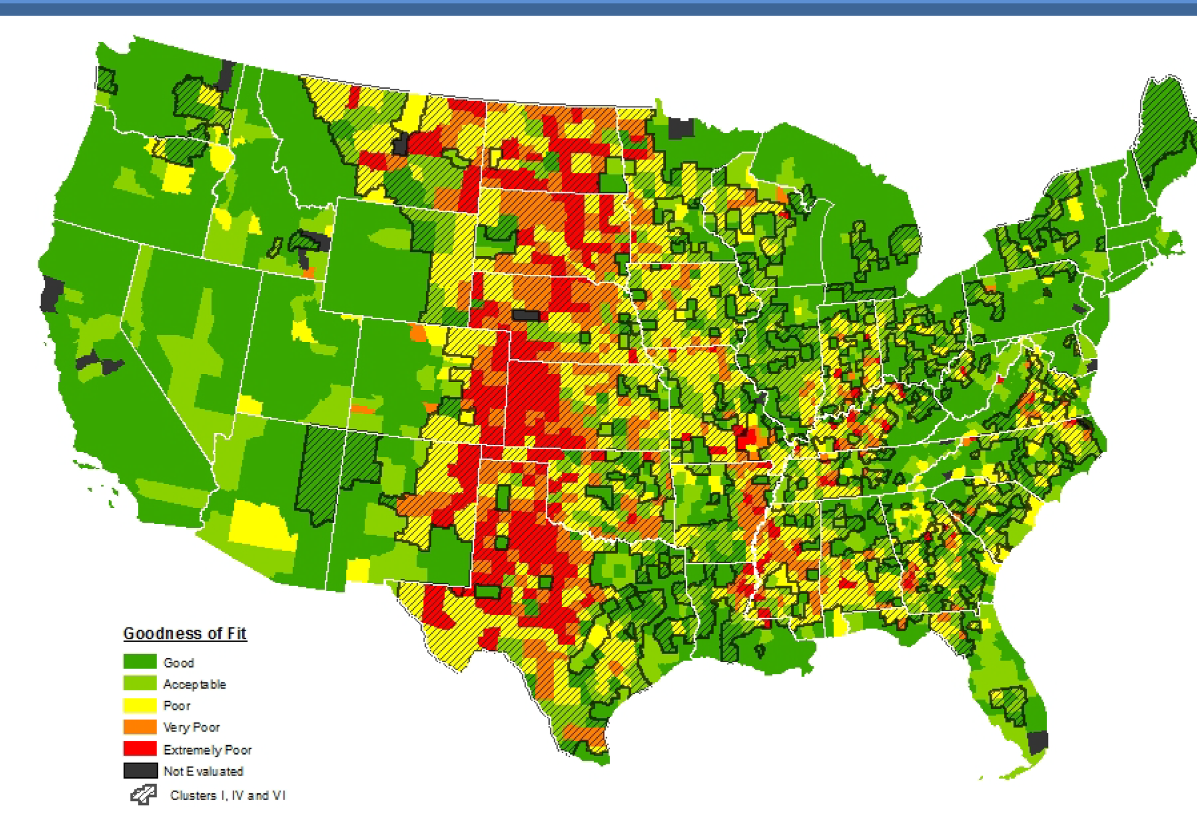
To ensure the correct performance of SWizard, we ran the script for an automatic calibration based on the 2001-2006 time period, followed by a prediction (option 6) with 25 Monte Carlo trials. This represents the most complex process in SWizard, so that most of its functionality for other options is tested. SWizard spent **19 hours, 23 minutes and 53 seconds** to make the calibration and prediction of the 3,109 counties, which is a considerable reduction in time if we were to make the scenario files and run SLEUTH manually.

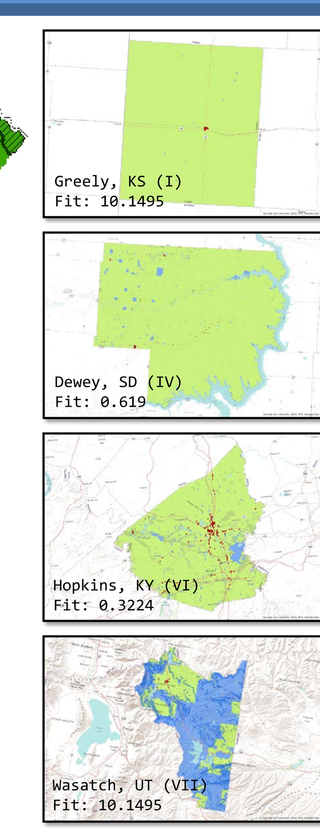
Table 3. Variables taken from the input data to characterize county's at the beginning o=f the period

Variable	Description
Urb01	% of pixels coded as urban at the beginning of the period (2001)
Available	% county percentage of suitable pixels for new urbanization in 2001. It is considered available those pixels that are not urbanized, not completely excluded and slope below 20%
AGr	Amount of available area in 2001 referenced to the growth observed
AUr	Amount of available area referenced to the urbanized area in 2001
Slope	Average slope in percentage of the available land in 2001
Patches	Number of contiguous clusters of available land in 2001
Patch size	Median area of available land patches in 2001
LAP	Percentage over the total available land of the largest patch in 2001
Centers	Number of contiguous urban clusters in 2001 greater than 3 pixels
LC	Percentage of urban area in 2001 that belongs to the largest urban center

In spite of this, there were some cases where SLEUTH had trouble running, in 21 counties of 3,109; we found that several of those cases were related to layers that present a road cover of 100% in the county. In the cases where SLEUTH crashes, SWizard just skips that county and registers the issue.

Results





Counties with a large amount of land available to urbanize tend to demonstrate low model fit; this tendency is aggravated when the ratio between available land and urbanized land is very high. Fragmentation of available land and slope are factors to reduce differences between model and observation. Counties with similar available area, if this is fragmented or less flat, usually have better scores because of the effect that those factors have in narrowing the suitability for development. Groups I, IV, and VI, with the worst results, clearly show a correlation with the degree of urbanization, and fit scores increase when the amount of urban land increases. The same occurs with the number of urban centers (urban clusters with more than 3 pixels), but at a much more reduced scale. Since the metric is a combination of measurements based on urban area, urban edge, and number of clusters, it is logical that those factors affect the level of accuracy in the metric. Meanwhile, it seems that the urban effect is not as strong in groups II, III, V, and VII, and probably the inaccuracy in clusters or edges has more relevance.

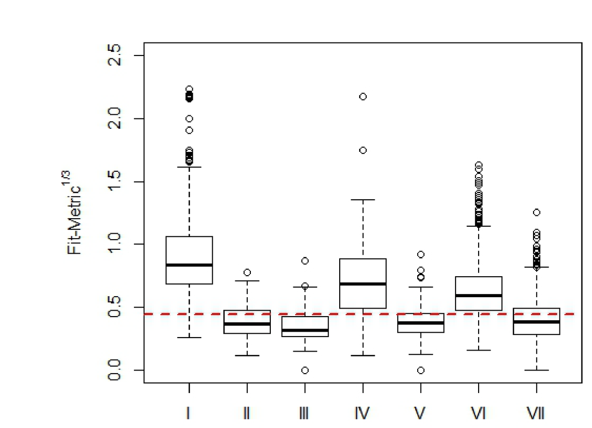


Figure 5. Box-plot of Global Fit Metric for 7 k-mean groups. Red line represent the threshold of 0.086.

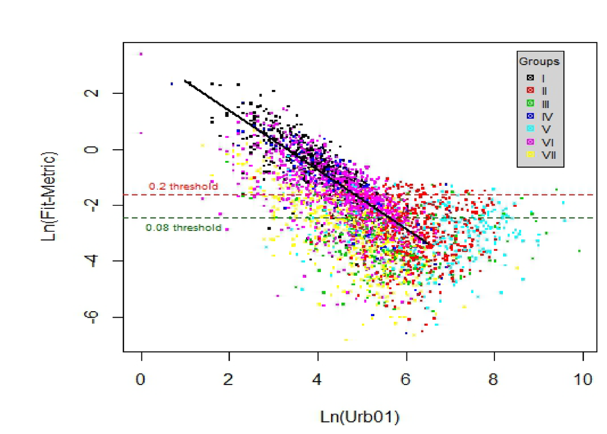


Figure 6. The accuracy of the model shows a dependency with the urban extension in groups I, IV, and VI

To be considered an acceptable goodness of fit, all the statistics involved in the metric must be within ±5% of the 2011 values (Jantz at al. 2010; Jantz, Drzyzga, and Maret 2014); however, only 40% of the predictions met this requirement (scores below 0.086.) The goodness of fit shows a clear spatial pattern distribution (see map avobe), where counties with a low model accuracy tend to be concentrated in specific regions. Using k-means, we identified seven groups that correspond reasonably with the distribution of the poorer scores. Groups I, IV, and VI clearly exhibit much higher values than the rest of groups do (fig. 5). The analysis of the groups determines there are statistical differences between groups, and the groups I, IV, and VI have higher values on average than the critical value of 0.086.

Table 4. Groups' mean value of variables described in table 3 and ordered by descending global fit metric score. The amount of available land for new developments appears as the main factor for the model's accuracy. The fragmentation of the available land and a less flat topography increases SLEUTH's reliability.

Group	Urb01	Available	Slope	Patches	Patch size	LAP	Centers	LC	AGr	AUr	Global Fit
I	0.43	98.16	1.42	1.39	11328.51	99.99	3.98	29.42	12,390.87	478.87	1.16 **
IV	0.30	84.11	2.62	73.82	3.03	97.37	7.49	30.40	31,842.43	815.71	0.60 **
VI	1.03	93.05	1.51	11.70	60.46	99.39	7.46	24.33	7,800.21	183.08	0.39 **
VII	1.42	65.73	5.73	80.12	2.54	85.77	9.39	19.26	4,949.66	111.62	0.11*
II	4.80	84.22	1.69	52.08	1.15	97.53	31.73	26.93	605.72	26.80	0.07
V	15.06	44.38	2.01	176.45	1.44	60.59	48.52	46.10	199.05	5.40	0.07
III	0.87	31.74	5.13	427.80	2.63	60.76	21.16	24.73	9,249.66	147.06	0.06

** Significantly greater than 0.2 (P<0.0005)

• Significantly greater than 0.086 (P<0.001)

Conclusions

- We demonstrated that SWizard is a useful tool for saving time for those who are using SLEUTH to model land use change, especially when users have multiple study areas or sub-regions to model.
- This script opens the door to model extensive areas and for studying differences across a large number of regions. It also provides improved capacity for testing the adjustment and behavior of the model, due to the amount of results that can be generated in an easier way.
- The analysis of 3,090 counties allowed for the identification of some factors that could affect model accuracy. Certain combinations of land available to new urbanization, urban growth, urban pattern, and topography have high likelihoods of producing poor fit statistics. In general, we can say that the more "freedom" the model has, the more inaccuracy it shows. That brings up the importance of the *exclusion* or *exclusion-attraction layer* (Jantz et al. 2010).

References

Dietzel, C. and K.C. Clarke. 2007. Toward Optimal Calibration of the SLEUTH Land Use Change Model. *Transactions in GIS* 11(1): 29-45

Jantz, C. C., and L. Morlock. 2011. Modeling Urban Land Use in the Upper Delaware River Basin

Jantz, C. A., S. Drzyzga, and M. Maret. 2014. Calibrating and Validating a Simulation Model to Identify Drivers of Urban Land Cover Change in the Baltimore, MD Metropolitan Region

Jantz, C. A., S. J. Goetz, D. Donato, P. Claggett. 2010. Designing and implementing a regional urban modeling system using SLEUTH cellular urban model.

Santé, I., A. M. García, D. Miranda and R. Crecente. 2010. Cellular automata models for the simulation of real-world urban processes: A review and analysis

Clarke, K. C., and Gaydos, L. J. 1997 Loose-coupling a cellular automaton model and GIS: long-term urban growth prediction for San Francisco and Washington/Baltimore, International Journal of Geographical Information Science, 12(7), 699-714

Silva, E.A., and Clarke, K. C. 2002 Calibration of the SLEUTH urban growth model for Lisbon and Porto, Portugal, Computers, Environment and Urban Systems, 26, 525–552

Clarke, K. C, Gazulis, N, Dietzel, C. K. and Goldstein, N. C. 2007 A decade of SLEUTHing: Lessons learned from applications of a cellular automaton land use change model. Chapter 16 in Fisher, P. (ed) Classics from IJGIS. Twenty Years of the International Journal of Geographical Information Systems and Science. Taylor and Francis, CRC. Boca Raton, FL. pp. 413-425.

Funding for this project comes from the William Penn Foundation. The William Penn Foundation, founded in 1945 by Otto and Phoebe Haas, is dedicated to improving the quality of life in the Greater Philadelphia region through efforts that increase educational opportunities for children from low-income families, ensure a sustainable environment, foster creativity that enhances civic life, and advance philanthropy in the Philadelphia region. In partnership with others, the Foundation works to advance opportunity, ensure sustainability, and enable effective solutions. Since inception, the Foundation has made nearly 10,000 grants totaling over $1.6 billion. The Foundation's assets exceed $2.3 billion as of Nov. 30, 2014. More information about the foundation is available on its website at www.williampennfoundation.org.

Mapping and monitoring of wetland dynamics for improved resilience and delivery of ecosystem services in the Mid-Atlantic region NNH10ZDA001N-LCLUC

2013 Summer at Census Program. Estimating the Completeness of TIGER Road Data Using the SLEUTH Urban Growth Forecasting Model