A Hybrid Approach to Population Construction for Agricultural Agent-Based Simulation

Peng Chen, Eduardo Izquierdo and **Beth Plale**

School of Informatics and Computing

Tom Evans

Dept of Geography

Michael Frisby

Indiana Statistical Consulting Center

Indiana University, Bloomington, Indiana USA

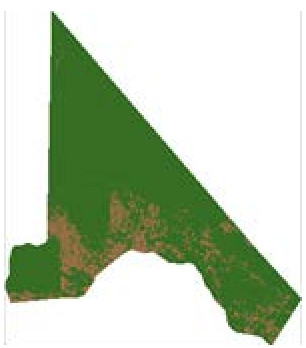
DATA TO INSIGHT CENTER
INDIANA UNIVERSITY
Pervasive Technology Institute

Introduction

- The advent of widespread fast computing has enabled us to work on more complex problems and to build and analyze more complex models.
- Agent-based modeling (ABM) is a key method in computational science. ABM is applicable to complex systems embedded in natural, social, and engineered contexts, across domains that range from engineering to ecology
- Spatial agent-based modeling (ABM) has been proven to be beneficial to agricultural economics for its ability to represent interactions amongst heterogeneous actors.

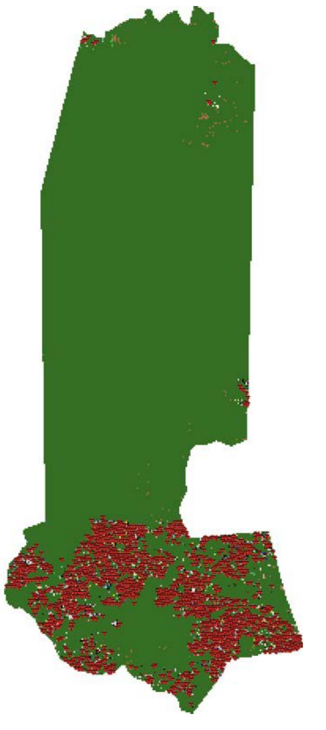
Motivation

- Agricultural economics researchers study ways in which humans can sustain themselves while not depleting an ecological/environmental resource
- When applied to small farms and individual farmers especially in countries such as Africa, a key element to harvest success is labor sharing
- It has been observed that farmers will share family members (labor) with neighbors and neighboring villages under certain circumstances



Motivation

- Agricultural economists build and analyze more complex models to understand labor sharing behavior
- Spatial agent-based models (ABM) have proven beneficial to agricultural economics for its ability to represent interactions amongst heterogeneous actors, and to fully take into account spatial dimension of agricultural activities



Agent Based Model (ABM)

- Zambia Agent-Based Model (ABM)

Agent Attributes

Household Agent

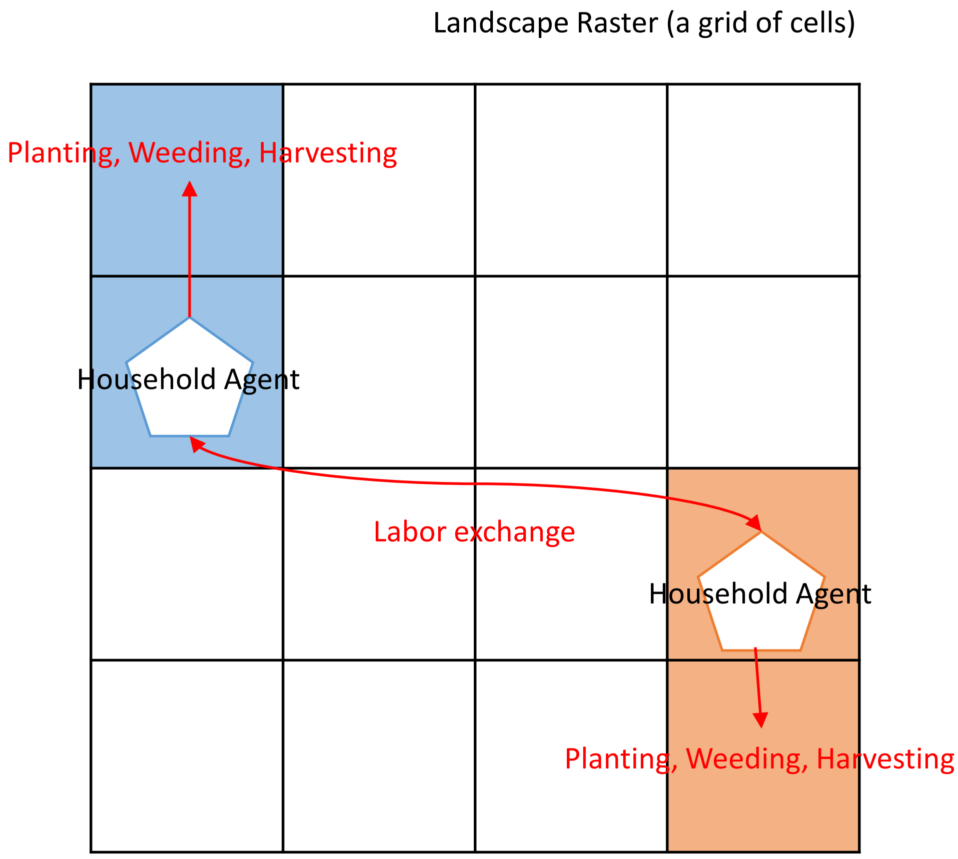
Number of household members (HH Size)
Area of cultivated land (CultArea)
Amount of labor
Amount of food stock
Amount of asset
When to plant
...

Agent Activities

Planting	Weeding	Labor exchange	Harvesting	...

Monze District, Zambia
53,491 households
1,866 square miles

Agent Based Modeling (ABM) Cont.



Agent Spatial Interactions

Left: agricultural land (brown) and non-agricultural land (green);

Right: households (red) allocated to agricultural land.

ABM challenge: configuring agents

- Agent-based models (ABMs) are highly sensitive to definition of the agents: their granularity, distribution, etc.
- Key to good agricultural agent based modeling is to construct agents that can truly reflect characteristics of real population of ***households***
- However, real population data about farmers and farming in Zambia is scarce
  - Limited
  - Insufficient
  - Aggregated
  - Not at a household level

Our Solution

- A hybrid approach to population construction

Do we have the agent variables in real population data?	Our Solution	Contribution
Yes	Simulating synthetic population data based on available datasets	Simulated data can have the same variability and heterogeneities
No	Calibrate missing variables with Genetic Algorithms (GAs)	1. Derived variables are optimized for replicative validity of the model. 2. We implement an microbial genetic algorithm that can: 1) Evaluate the fitness based on the behaviors of all agents; 2) Handle the stochasticity in the simulation run.

Related Work

- Creation of household agents in ABMs: agricultural analysis (Evans, 2004) (Kelly, 2011), urban planning (Beckman, 1996) and urban disaster management (Felsenstein, 2014).
  - focused on decomposing aggregated demographic/administrative data
- Environmental modeling: create agents from survey data (e.g., parameterisation) (Iwamura, 2014) and agent typology (Valbuena, 200).
  - None integrate real population data into agent creation process
- Genetic Algorithms (GAs): automatically search a parameter space, and thus they have been used to calibrate agent-based models (Calvez0, 2005),(Espinosa, 2008), (Wu, 2002), (Mulligan, 1998).
  - Challenges remain in how to design fitness function that can consider behaviors of all agents; and stochasticity in simulation run.

Outline

- Introduction
- Related Work
- Proposed Hybrid Method
  - Simulation of Synthetic Population
  - Calibrating Agent Variables with GA
- Application and Evaluation
  - Zambia Food Security ABM
  - Household Characteristics Simulation
  - Variables Calibrated by Microbial GA
  - Summary

Real Data Sources for Population Data

- Farmer Register
  - Small scale farmers, total area under cultivation
  - 53,579 records
- Household Survey data
  - Compiled by regional agricultural extension officers
  - Census of all small-scale farmers in particular district
  - Basic attributes: total area of farm, total area under cultivation in particular year
  - 330 households

Real Data Sources for Population Data

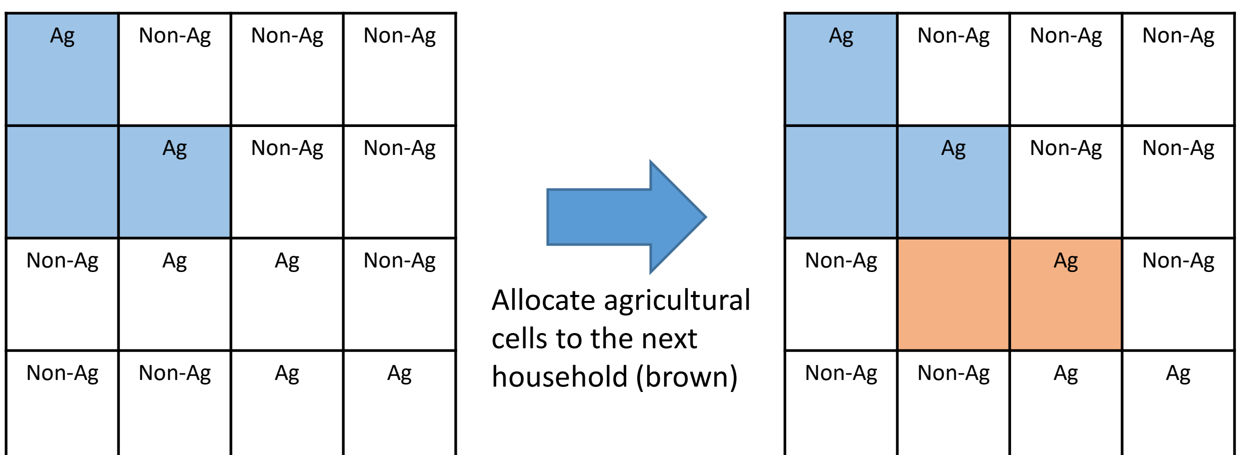
- Post Harvest Survey data
  - Used by Zambian government to assess crop yield
- Remote Sensing data
  - Classifies gridded images into agricultural and non-agricultural land
  - Disaggregates features to raster (vector) data form
- Need: develop land allocation algorithm that can form natural farmer communities when placing the household agents

Recall

- From known data from multiple sources (all spotty) , get good starting set of agents as households that farm land of known (and representative size). Households of representative wealth, # household members, etc.
- Fill in critical missing data using Microbial Genetic Algorithm:
  - soil type,
  - ratio of hybrid maize to local maize planted,
  - planting data standard deviation

Simulating Household Spatial Locations

- Input remote sensing data
  - Classified and disaggregated into agricultural and non-agricultural land cells
- Our land allocation algorithm then allocates the agricultural cells to households
  - First chooses a number of seed households and randomly assign agricultural cells to them.
  - Then each time assigns to a household with an unallocated agricultural cell that is adjacent to some allocated agricultural cell.



Calibrating Agent Variables with GA

- Genetic Algorithm (GA): heuristic search that mimics process of natural selection:
  - Start with population of individuals and fitness function
  - Properties of individuals are mutated and altered in each generation
  - Best fitted individuals are preserved to next generation
- Microbial Genetic Algorithm is minimal GA that has same functionality and efficacy as standard Gas
- Most creative and challenging parts of programming a GA are:
  - Chromosome – set of properties for each individual in population – and its mutation/alternation process
  - Fitness function – fitness score is usually objective value in optimization problem being solved

Calibrating Agent Variables with GA Cont.

- Chromosome could be composed of properties that each represents a missing agent variable:

Table: different types of properties in a chromosome

Type	Example	Representation
Nominal variables	soilType	Represented as an integer that can be randomly mutated into any other possible values
Simple continuous variables	ratioOfLocalMaize	Represented as doubles, and can be mutated with a Gaussian number generator.
Variables that follow a certain distribution	plantingDate that follows a normal distribution	Represented as a parameterized distribution, whose parameters can be mutated with a Gaussian number generator

Calibrating Agent Variables with GA Cont.

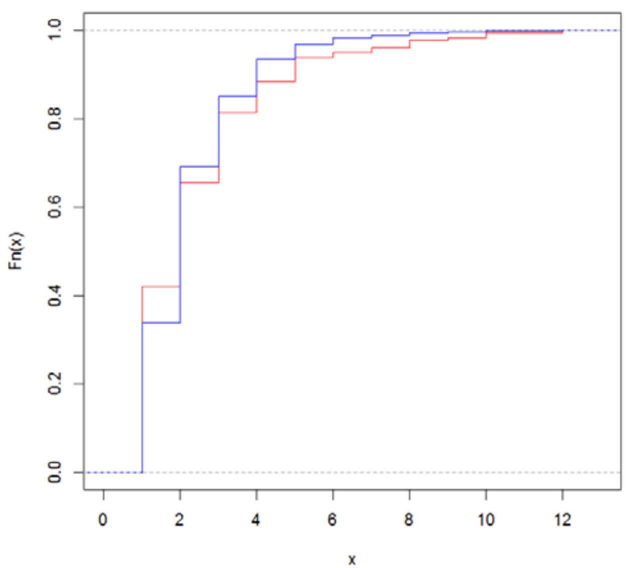
- We use distance between simulated outcome and real world observations as fitness score
  - Data generated from agent-based model can be collected at individual level (e.g., yield of each household agent) or at aggregated level (e.g., total crop production). Model calibration needs to be at both levels.
  - We use Kullback–Leibler divergence to measure difference between distribution of simulated data and distribution of observed data
- ABM is stochastic in that two simulation runs can produce different results
  - We explicitly set random number seed (R) in agent-based model and expose R as property of GA chromosome to handle stochasticity

Outline

- Introduction
- Related Work
- Proposed Hybrid Method
  - Simulation of Synthetic Population
  - Calibrating Agent Variables with GA
- Application and Evaluation
  - Zambia Food Security ABM
  - Household Characteristics Simulation
  - Variables Calibrated by Microbial GA
  - Summary

Zambia Food Security ABM

- ABM of agricultural decision-making on Monze District, Zambia
- Clean survey data and Farmer Register
  - Extract from huge spreadsheet
  - Round Cultivated Area (CultArea) to integers
  - Remove incorrect values and outliers
- Classify and disaggregate remote sensing data



After cleaning, survey and Farmer Register have similar Empirical Cumulative Distribution Functions (ECDFs) for rounded CultArea

Red: rounded variable CultArea from survey data
Blue: rounded variable of CultArea from register data

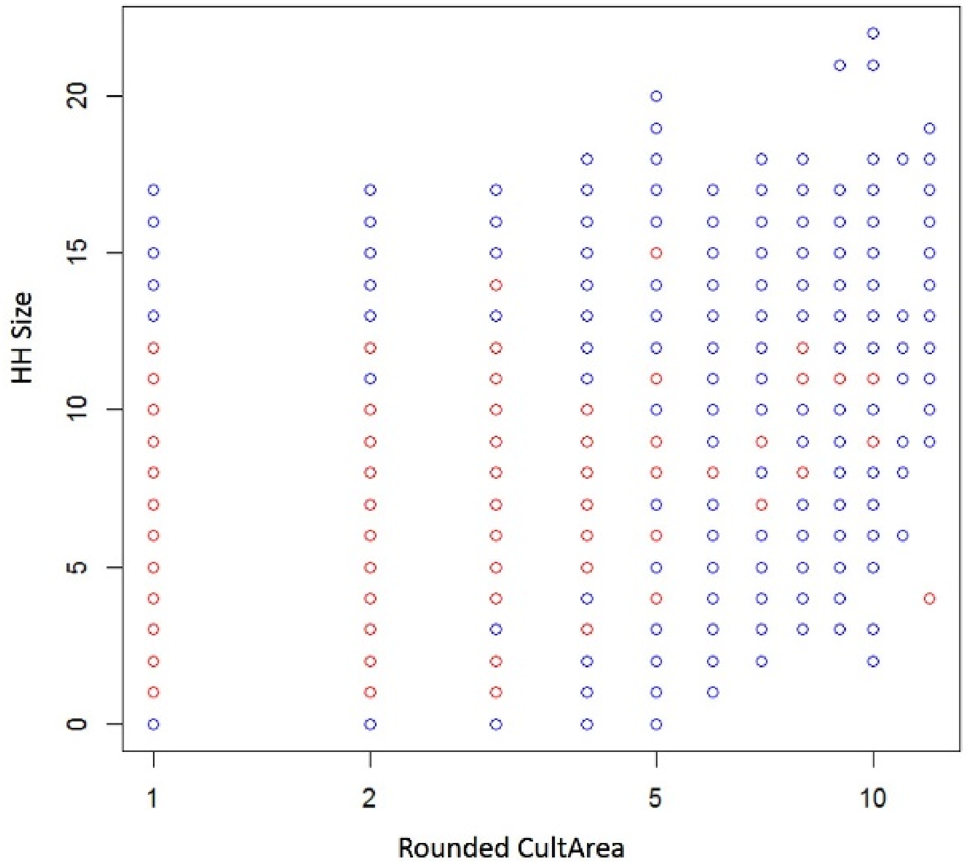
Household Characteristics Simulation

- Independent variable X – *HHSize*
  - The household size (i.e., the number of members in a household) is modeled with a Poisson distribution.
- Dependent variable Y – *CultArea*
  - Cultivated variable is missing in farmer register
- We fit a generalized linear model with the variable *CultArea* (rounded) and *HHSize* from the survey data.

$$\log(E(HHSize \mid CultArea)) = a + b * CultArea$$

- We use fitted model to predict mean value of *HHSize* for each value of *CultArea* in farmer register.
- Finally we use predicted mean value of Poisson distribution to randomly generate simulated values of *CultArea*

Household Characteristics Simulation Cont.



Distribution of cultivation area per household size; overlaying simulated data (blue) and survey data (red). X-axis is in log scale.

Household Spatial Location Simulation

Results of land allocation in one ward of Monze District, Zambia.

Left: agricultural land (brown) and non-agricultural land (green);

Right: agricultural
Land allocated to households (red).

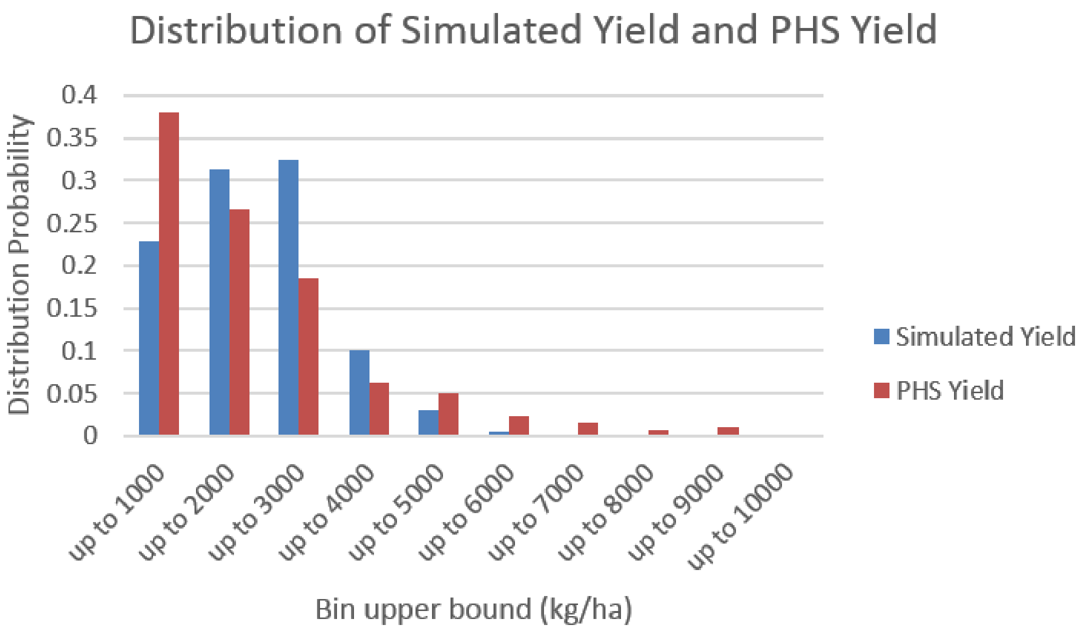
Variables Calibrated by Microbial GA

- Finally, calibrate all missing variables whose values could not be determined in previous steps
- Each chromosome is composed of four properties:
  - soilType – integer: [0, 14]
  - ratioOfLocalMaize – double: [0, 1]
  - plantingDateStandardDeviation – double: [0.001, 0.167], which represents the standard deviation of normal distribution of planting date.
  - randomSeed – any integer

Discussion of use of Genetic Algorithm

- ABMs have lots of parameters that together determine global dynamics of model. Huge search space. GA's good at dealing with large dimensionality
- No mathematical equation that can anticipate dynamics of agent-based model without executing it, thus high computation load to determine fitness function (which requires repeated execution of simulation)
- Genetic algorithm is more efficient than Monte Carlo experiment
- Determination of predictive ability of GA is open question

Summary



Comparison between simulated yield and observed yield distribution from Post Harvest Survey

Next step: use synthetic population as basis for studying household interaction under different scenarios of climate change

Q&A

- Thanks!
- Contacts
  - plale@indiana.edu
  - chenpeng@indiana.edu
  - Data to Insight Center, http://d2i.indiana.edu/

DATA TO INSIGHT CENTER
INDIANA UNIVERSITY
Pervasive Technology Institute