Uncertainty and Visualization Issues of Microsimulation for Social-Cultural Modeling

Charles Ehlschlaeger: ERDC-IL; University of Illinois
Noon, 03 May 2010
Phelps Hall 3512
University of California at Santa Barbara
Abstract

Social-cultural behavioral modeling is increasingly seen as a useful tool to understanding complex behavior in unfamiliar cultures. During disaster relief and infrastructure improvement missions, non-governmental organization, USAID, NATO, and other organizations are working in foreign environments causing massive changes to existing social structures.
Abstract, cont.

...This research explores the data and tools being developed to better understand the impacts in these operations, especially the Digital Populations technique. Digital Populations generates multiple representations of all households and people in a geographic area, allowing more intuitive social-cultural models to be constructed. Several models will be shown.
Outline

• What’s so special about Social-Cultural Knowledge?
• Modeling Overseas S-C problems
• Digital Populations
• Modeling w/ Great Data
• Modeling w/ Good Data
• Modeling w/ Poor Data
• Future research plans
Social-Cultural Knowledge

- Ephemeral
- Hard to quantify
- Difficult to visualize by outsiders
- Important knowledge is in processes, not measurable information
- System components less precisely defined compared to environmental models
- Knowledge often represented within models
- Necessary knowledge for even simple problems requires multiple subject matter experts (even ignoring the programming geek)
Background on DoD S-C Modeling

- Typical S-C models require years to build
- Calibration and Validation often absent
- Dynamic environments often have `shocks’ that should modify model
- Thus
  - most models obsolete before finished
  - S-C needs `30 day models’ to be effective
  - ERDC-IL assisting with `30 day’ modeling efforts in non-spatial temporal modeling environments
Digital Populations

• DP is one piece of potential solution for Rapid-or Mediated Modeling approach to bring space-time to S-C behavior modeling

• Goal: To build representation of every man, woman, and child in study area containing `rich contextual knowledge’ about each person
Digital Populations (US States) Methodology

- Building Realizations of Digital Population
  - Modified National Land Cover Dataset (NLCD): 30 meter resolution data.
  - Grid cells subdivided into “close to water” and “normal” due to significant positive population density (great for Rhode Island, no effect for IL or Chicago)
  - American Community Survey (ACS): PUMS-like data on an annual basis.
  - U.S. Census Short Form (SF) aggregated data.
Building Realizations of Population

- Relative Household Density of modified NLCD classes (heterogeneous Poisson process):
  - Multiple step-wise regression: \[ h_i = \sum_{k} d_k c_{ik} + e_i \]
    
    \( h_i \) is number households in SF area \( i \).
    \( d_k \) is household density in NLCD class \( k \).
    \( c_{ik} \) is area of NLCD class \( k \) in SF area \( i \).
    \( e_i \) is error of SF area \( i \).
  - Iterative process: remove NLCD classes with negative density and repeat until all \( d_k \) is positive.

- (Improvements to this 1\(^{st}\) order process beginning this summer)
Building Realizations of Population

• Populate Study Area with ACS Households using Relative Household Density:
  – Census Areas chosen by two sets of criteria:
    • SF household occupancy and SF population
    • “Application specific” SF & ACS variables
  – Location within Census Area conditional stochastic process based on Relative Household Density
  – Why application specific variables, not all?
    • Experimental results with Digital Populations and other MCS processes (Ehlschlaeger 2002) indicate that increasing number of variables to fit will decrease quality of individual variables
    • Great census data should model all variables
    • Good to poor data requires fewer variables for proper fit
Building Realizations of Population

- Once study area is initially populated, random households are relocated to new locations if variable fits improve
  - If cases available, process is conditional
    - Households with member(s) of target sub-population that are considered positive cases are fixed
  - This process is time consuming
    - 250 realizations of RI older African American women took one month of computer time (1 gig. RAM, 3.2 MHz Pentium IV)
      - Algorithm designed to allow different computers to compute realizations with repeatable results. (Easy to do in Java.)
Building Digital Populations: Theoretical benefits

- DP points better than Short Form data?
  - Easier to aggregate to any choropleth scheme
  - Easier to retain level of measurement when applying to point based applications
  - Easier to retain uncertainty information
Digital Populations Methodology: land cover and census areas
Digital Populations Methodology: One realization of possible household locations
Digital Populations Methodology: One realization of sub-population locations
Digital Populations experimental results

- In limited experiments, DP and modified Kuldorff spatial scan statistic identifies simulated cancer clusters better than SaTScan and choropleth data.
- DP provides more accurate representation of population uncertainty:
  - Choropleth data treats population as living in centroid of each census block, tract, or county.
  - DP simulates exact household locations accounting for stochastic distribution across land use classes.
Modeling with Great Data & Knowledge
Modeling with Great Data & Knowledge

- Modeling Domestic Violence in Chicago
- Marina Drigo’s Masters Thesis @ UIUC
- Geographic data locating social health centers assisting victims of domestic violence
- Extensive literature review mining statistical representation of actors’ actions
- Digital Populations representation of households adds neighborhood level spatial accuracy
Model with Great Data
`What If’ modeling

Simulation duration: 5 Years, 9 Months

Option to change community center capacity
- Percentage of center capacity (%)
- Center status
- Original capacity

Option to change shelter capacity
- Percentage of shelter capacity (%)
- Shelter status
- Original capacity

Option to change public awareness
- Change awareness
- On/Off

Option to include extra center services or shelters (by number and location)
- Add extra services?
- # of services to add
- Area for new service

Percent victims
- 0.44%
- Red: White
- Green: Hispanic
- Blue: Black

Percent of population
- 0.0%
- Months
- 73.8

Real rate per 1,000 couples
- 26.7
- Months
- 73.8

US Army Corps of Engineers
Engineer Research and Development Center
Model with good data
Modeling Overseas

- Food Security for Cartagena Colombia
- IPUMS for district containing Cartagena
  - (Lucky of us, Colombian census didn’t collect for all of district, just Cartagena)
  - 2005
- Census Data for Cartagena
- Landcover for Cartagena
- Food Security Model
Economics for Cartagena Model

- “Most” food initially distributed at one Mercado
- Lower income people in Cartagena purchase “most” food at tiendas
- Tiendas act as convenience stores, restaurants, and informal banks
- Half of Cartagena population works in informal jobs
GIS data - IPUMS

- International Public Use Microdata Samples
- United Nations has published standards of attributes to be collected
- Individual nations often choose subset of attributes
- Contains Household and Personal Information
IPUMS Household variables may include:

- Technical information (metadata about household)
- Group quarters (# unrelated people)
- Geography (household in urban/rural, region, department, metro area, municipality recode, & head town)
- Economic information (ownership, international migrants)
- Utilities
- Appliances, Mechanicals, & other Amenities
- Dwelling Characteristics
- Constructed Household (# of families, couples, mothers in household)
IPUMS Person information may include:

- Technical (metadata about record)
- Family Interrelationship
- Core Demographic (age, sex, marital status)
- Fertility and Mortality (information about children)
- Nativity and Birthplace (where born, year of immigration)
- Ethnicity and Language
- Education
- Work
- Income
- Migration
- Disability
IPUMS

• Depending on goal, IPUMS variables can provide insight
• Sub-population representation critical for many applications (may be demonstrated at end of this discussion)
Cartagena Food Security Model

- Higher food prices prevents many people from buying higher priced protein rich foods
- Modeling food distribution from Mercado to tienda to household provides estimate of protein to households based on their income
- Modeling households as individuals with age & gender estimating food & protein per person
- User can model change in food prices, food aid, and is easily adjusted
Modeling Tools to aid Decisions

US Army Corps of Engineers

Engineer Research and Development Center
However...

- Data requires extensive massage
- Cultural differences increases likelihood of model misrepresentation (even with subject matter experts looking over model development)
- Model development in US for overseas area of interest decreases iterative calibration and validation opportunities
Modeling w/ Poor Data
Modeling Socio-Cultural Information in Afghanistan

- Afghanistan has poor history of demographic representation: last official census 1979
- Validation of recent demographic and other social cultural information difficult to impossible due to lack of historical data
- Current data collection often biased by short term needs
- Open source information has contradicting information
- Knowledge of demographic and socio-cultural data collection techniques not taught in AF universities
Landcover for Digital Populations

[Map showing various landcover types]
Cultivated Areas (agrarian society)
Roads
Settlements and Cities
Health Care Facilities
Discussion of Data Limitations

- Survey data for specific needs, and not likely valid for other needs
- Data collected by many sources with poor quality standards
- Web 2.0 data collection techniques not easily done
Next Generation of Digital Populations

- Households of sub-populations located in Google Earth environment
- Query tools to identify sub-population density
- Visualization of indicator variables
- Use of Digital Population data to determine utility of qualitative data
  - Measure sources of qualitative data
  - Determine whether qualitative data can be generalized to larger population
- Easily feeds into Socio-Cultural modeling environment
  - NetLogo
  - Repast
  - Cultural Geography Model
Rapid Modeling Environment for Socio-Cultural Knowledge Representation

• Research program to synch Digital Populations with agent-based socio-cultural models representing S-C behavior
  – Kickoff research effort
  – ERDC research plans includes eight year effort (in an elegant Gaussian shaped distribution)

• Goals:
  – Better situational awareness when improving infrastructure and essential services in areas with low demographic information
  – Better `what if’ planning in `wicked problem’ environments