Unpacking space-time dynamics from multi-time (non-longitudinal) spatial data

Li An

October 28, 2016

Department of Geography Colloquium San Diego State University

Acknowledgement

- The SDSU NASA project team: Doug Stow, John Weeks, Pete Coulter, Helena Taflin (Idea, data, and funding)
- Drs. Atsushi Nara, Andrew Zhang, and Ke Huang at SDSU (Python programming)
- Evan Casey at SDSU (Data preparation)
- Dr. Guangming He (programming and parallel computing)

Methodology background

- Many social, political, and demographic / health surveys are conducted multiple times
  - Type A: The same set of subjects are surveyed → longitudinal data
  - Type B: (In many instances) Each time a different set of subjects are surveyed → multi-time, non-longitudinal (spatial) data.

Type A data

- Women's Health Study of Accra
  - Wave I: 3200 (2003)
  - Wave II: 2814 women (2008/2009)
- Chitwan Valley Family Study (CVFS)
  - Mixed method longitudinal study in Chitwan Valley, Nepal
  - 10,000 individuals were followed in 1996, 2001, and 2006

Type B data

- Gallup surveys
  - Around 1,000 people were interviewed each year (for over two decades)
  - Each year a **different** sample is taken
- Ghana Demographic and Health Survey
  - (1988) 1993, 1998, 2003, 2008, and 2014
  - Each time a different set of HHs were surveyed

Methodology frontier

- Many readily useable methods for Type A data
  - Survival (hazard, event history) analysis
  - Latent trajectory modeling

Longitudinal data analysis

- **An, L.**, M. Tsou, B. Spitzberg, J.M. Gawron, and D.K. Gupta (2016). Latent trajectory models for space-time analysis: An application in deciphering spatial panel data. *Geographical Analysis* 48 (3): 314–336.
- Crook, S.E.S., **L. An**, D.A. Stow, and J.R. Weeks (2016). Latent trajectory modeling of spatiotemporal relationships between land cover and land use, socioeconomics, and obesity in Ghana. *Spatial Demography* 4(3): 221-244.
- **An, L.**, D. G. Brown, J. I. Nassauer, and B. Low (2011). Variations in development of exurban residential landscapes: timing, location, and driving forces. *Journal of Land Use Science* 6(1):13-32.
- **An, L.**, and D. G. Brown (2008). Survival analysis in land-change science: integrating with GIScience to address temporal complexities. *Annals of Association of American Geographers* 98(2):323-344.

Methodology challenge

- Lack of effective methods for Type B data
  - Aggregate to higher level (e.g., individual → city)
  - The curse of “ecological fallacy”

Research questions

1. How to derive spatiotemporal trends based on multiple-time, yet non-longitudinal data (Type B)?
2. How to find out the associated mechanism?

Case study

- The Urban Transition in Ghana and Its Relation to Land Cover and Land Use Change through Analysis of Multi-scale and Multi-temporal Satellite Image Data (PI: Doug Stow)
  - Interdisciplinary Research in Earth Science (IDS) program
  - NASA Award #: G00009708.
  - The four regions of the Greater Accra, Central, Eastern, and Ashanti

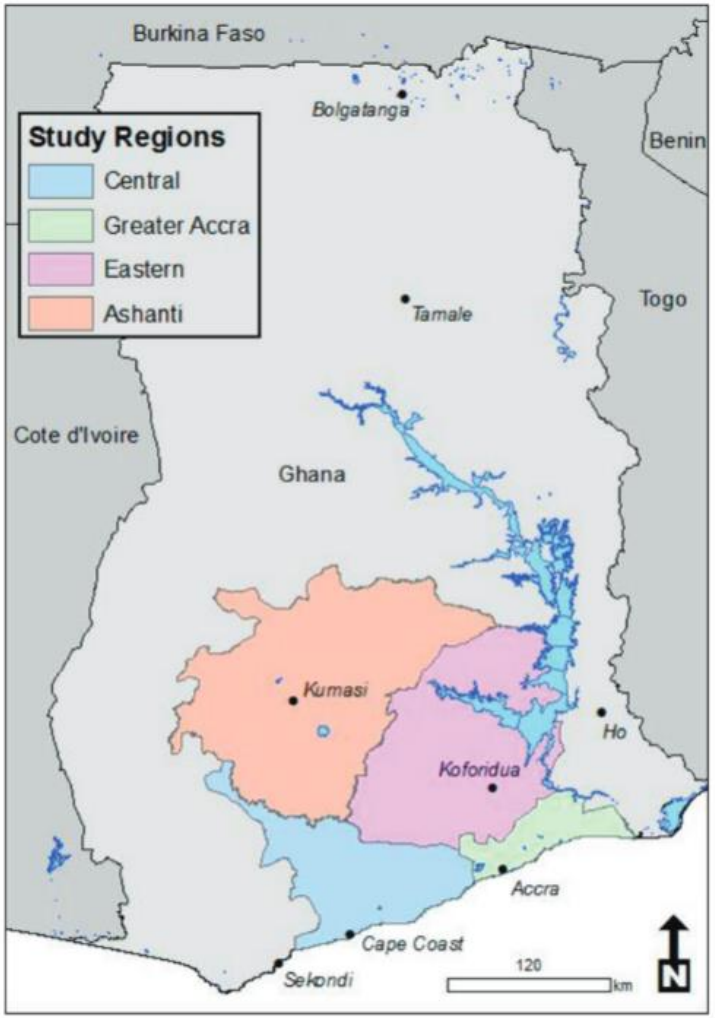


Fig. 1 Study area in Ghana

Background
(Nutrition transition)

(1) Hunter Gatherer

(2) Early Agriculture

(3) End of Famine

(4) Overeating and Obesity-related Diseases

(5) Behavior Change

Popkin (1993, 2002, 2009), Crook et al. (2016)

Body mass index

- BMI = A person's weight (kilograms) divided by the square of that person's height in meters (Kg/M^2)
- Standards
  - BMI < 18.5: underweight
  - BMI > 25: overweight
  - BMI > 30: obese

Research questions

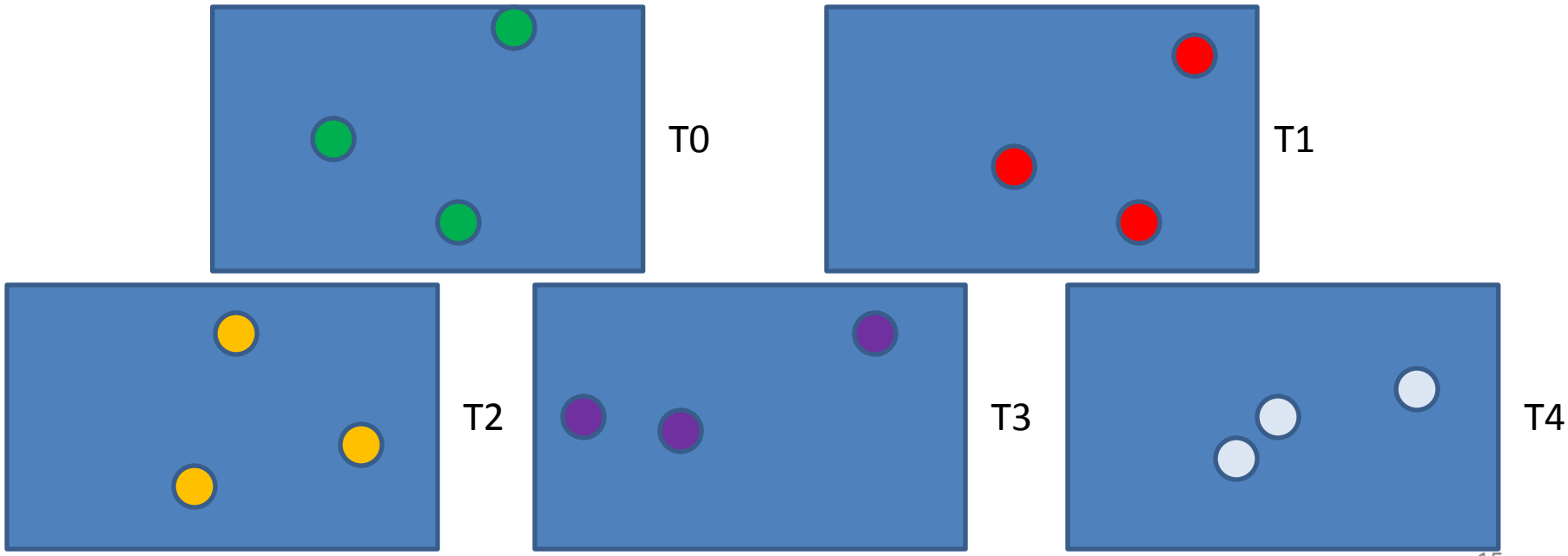
1. How to derive body mass index (BMI) spatiotemporal trends at individual level based on multiple-time, yet non-longitudinal data (Type B)?
2. How to find out the associated mechanism behind such trends?

Data

- Ghana Demographic Health Survey (DHS) data in 1993 (T0), 1998 (T1), 2003 (T2), 2008 (T3), and 2014 (T4)

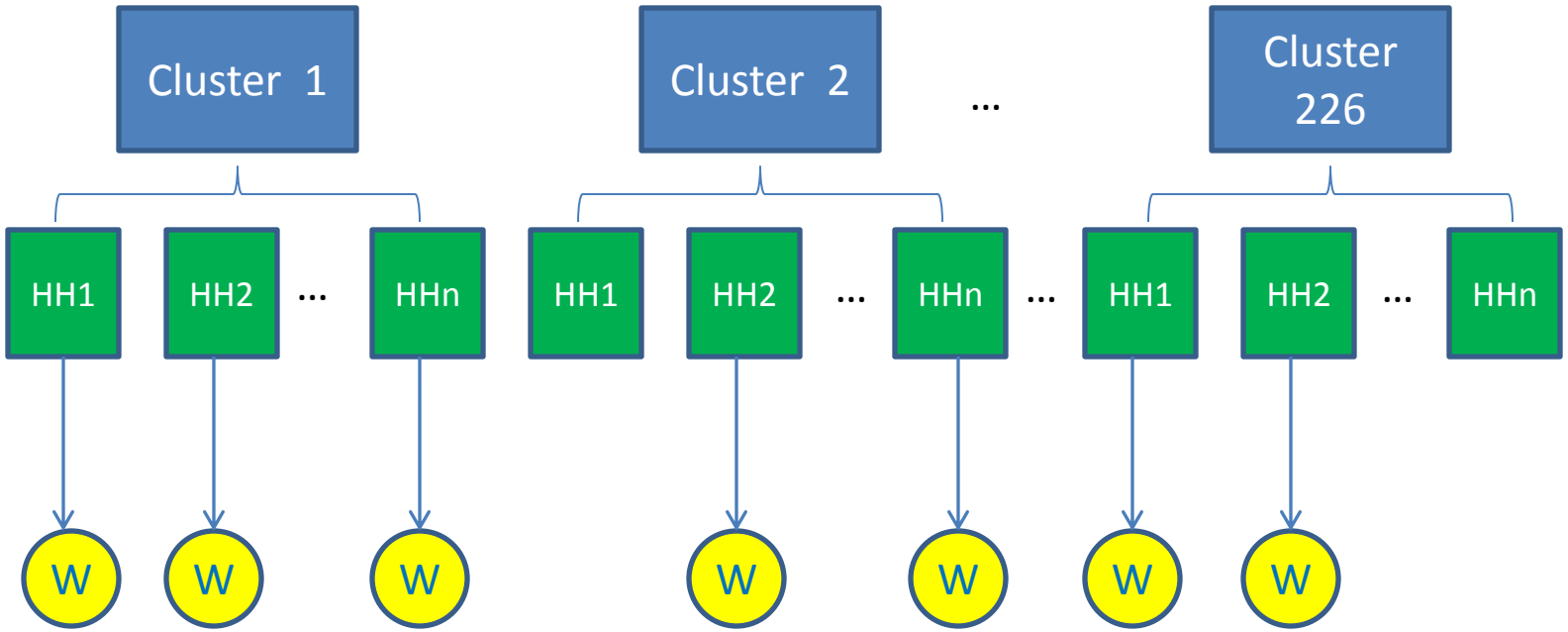
Data characteristics

- Five Ghana DHS samples, each time ~220 randomly clusters were sampled
- The sampled clusters (circles in the graphs below) vary from T0 (1993), T1 (1998), T2 (2003), T3 (2008), & T4 (2014)—They are Type B data
- Each cluster has a varying # of households;



Selection of subjects

- There are >200 clusters in 1998 DHS data



985 women are included in the sample

1. All HHs in the same cluster have to share ONE cluster centroid x and Y (data limitation);
2. Some HHs may not have eligible women → 226 women will be selected —but can expand to a larger sample

Table 1. Variable of interest

Variable name	Explanation	Scale	Temporal scale	Data source
BMI	BMI of each selected woman	Household (placed at cluster centroid) / cluster	Four times (~~1993,~~ 1998, 2003, 2008, and 2014)	DHS

Table 2. Independent Variables

Variable name	Explanation	Scale	Data source	Note	Who
UrbanLandCover[1]	See Crook paper	2.5 Km buffer from cluster ctr	Crook paper	Same as Crook paper	Crook paper; link by ID[2]
AgLandCover	See Crook paper	2.5 Km buffer from cluster ctr	Crook paper	Same as Crook paper	Crook paper; link by ID
NatVegLandCover	See Crook paper	2.5 Km buffer from cluster ctr	Crook paper	Same as Crook paper	Crook paper; link by ID
Top 10 filters	10 eigenvectors	Nearest 64 clusters	Crook paper	Same as Crook paper	Crook paper; link by ID
HH Size	See Crook paper	HH (cluster ctr)	DHS T0~T5	Same as Crook et al. 2016, but at HH level.	Evan will collect
FlushToilet	See Crook paper	HH (cluster ctr)	DHS T0~T5	Binary (not a %)	Evan will collect
NoToilet	See Crook paper	HH (cluster ctr)	DHS T0~T5	Binary (not a %)	Evan will collect
HasElectricity	See Crook paper	HH (cluster ctr)	DHS T0~T5	Binary (not a %)	Evan will collect
Age	See DHS meta-data	HH (cluster ctr)	DHS T0~T5	Collect (meet H & E)	Evan will collect
Wealth (?)	See DHS meta-data	HH (cluster ctr)	DHS T0~T5	Collect (meet H & E	Evan will collect
Sedentary	See DHS meta-data	HH (cluster ctr)	DHS T0~T5	Collect (meet H & E	Evan will collect
2nd-higher-edu	See DHS meta-data	HH (cluster ctr)	DHS T0~T5	Collect (meet H & E	Evan will collect
Partner@Home	See DHS meta-data	HH (cluster ctr)	DHS T0~T5	Collect (meet H & E	Evan will collect
(x, y) coord's	See DHS meta-data	HH (cluster ctr)	DHS T0~T5	For ABM only (Not used in regression)	Evan (Done)

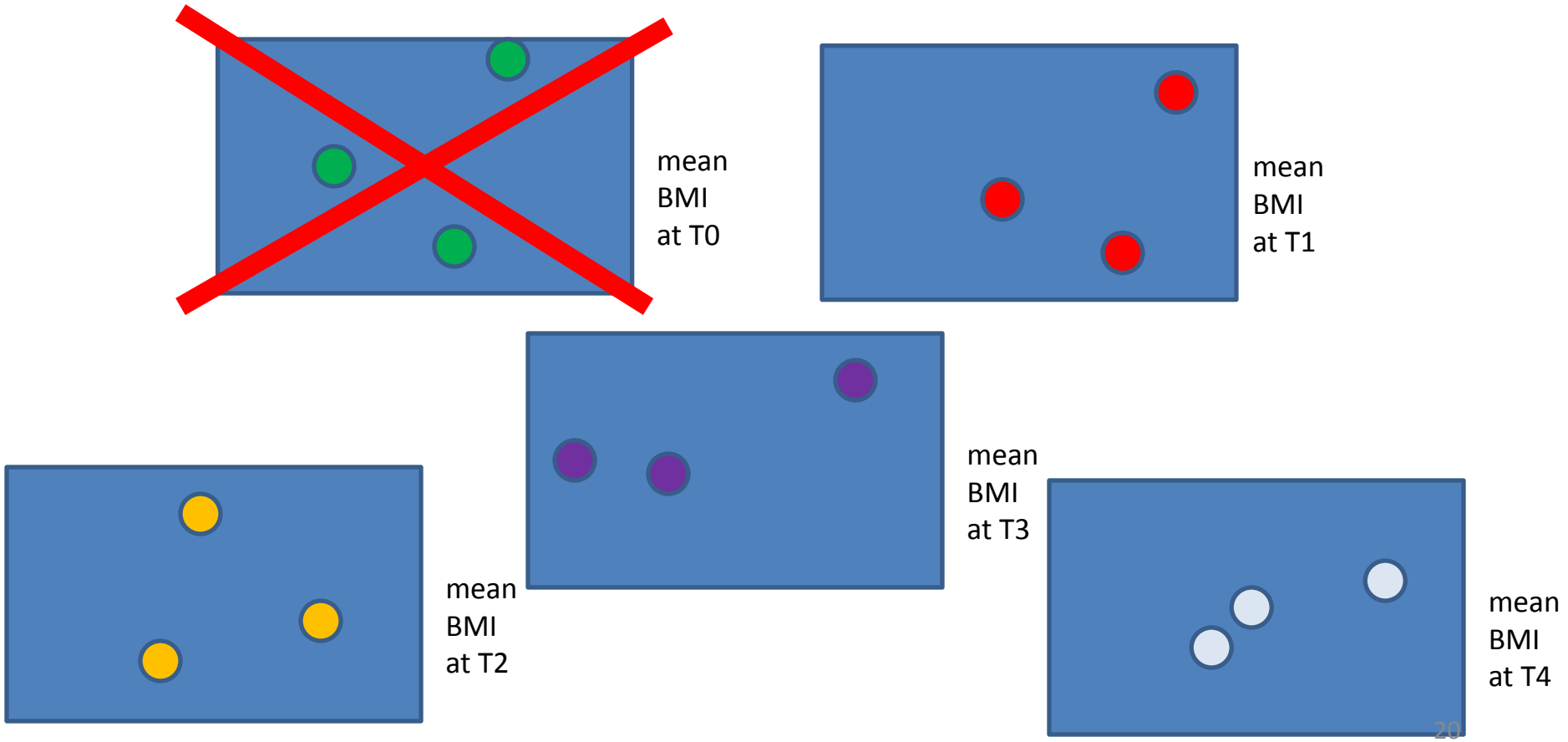
1. All variables in purple are the ones used in Crook et al. (2016), but collected at the same or HH level.

Derive spatiotemporal trends?

- Keep in mind we do not have longitudinal data, and cannot apply any longitudinal data analysis methods such as LTM
  - Subjects or units of analysis must be identical over time in longitudinal data analysis
  - Crook et al (2016) created longitudinal data through aggregation (kriging)
- Instead we have five random samples
  - In theory each sample should be very representative of the population at each time
  - If there is any change (based on the sample), it should be a change in the total population

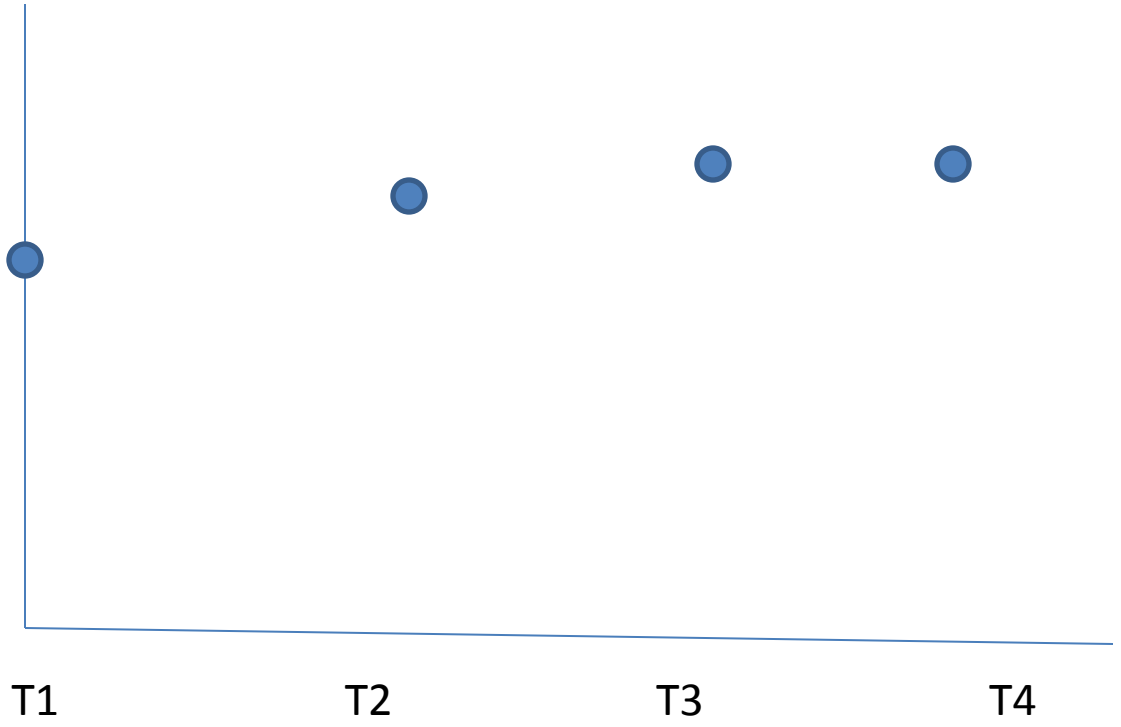
Step 1: Mean calculation

- Calculate mean BMIs T1 ~ T4



Step 2: Trend visualization

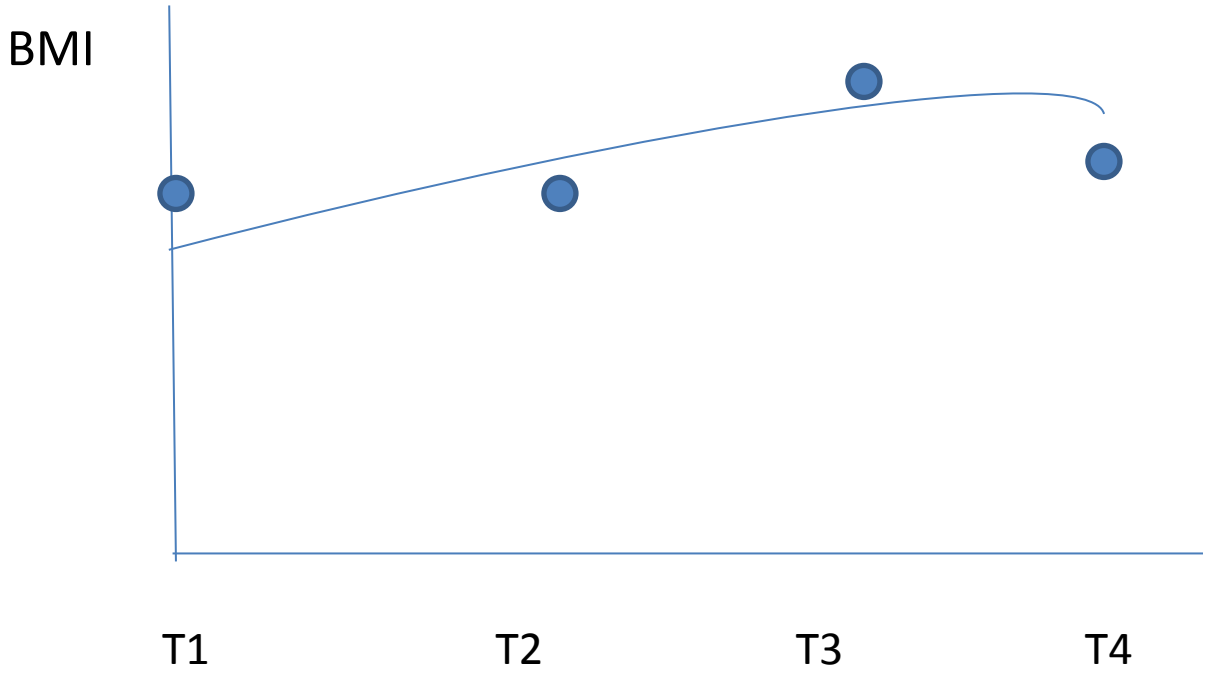
- Calculate and plot the five means over T1 ~ T4:



Step 3: Trend fitting

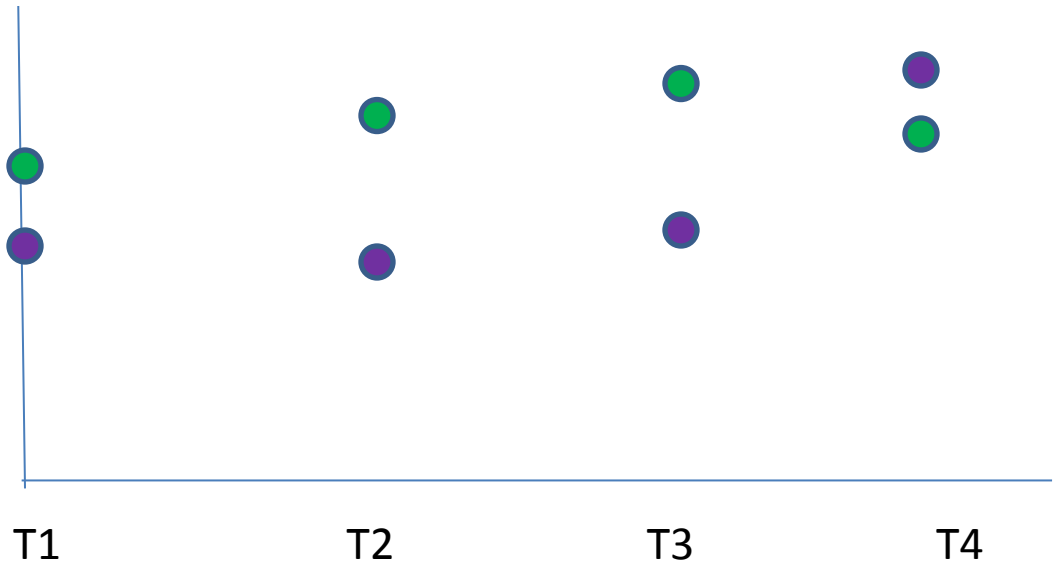
- Fit a quadratic line

$$BMI_t = \alpha + \beta_1 t + \beta_2 t^2 \qquad (1)$$



Create each individual's trajectory?

- For each individual at T1 (1998), we only know its BMI at T1—**BUT**:
- Can we still project its BMI in later years? Possibly **YES!**



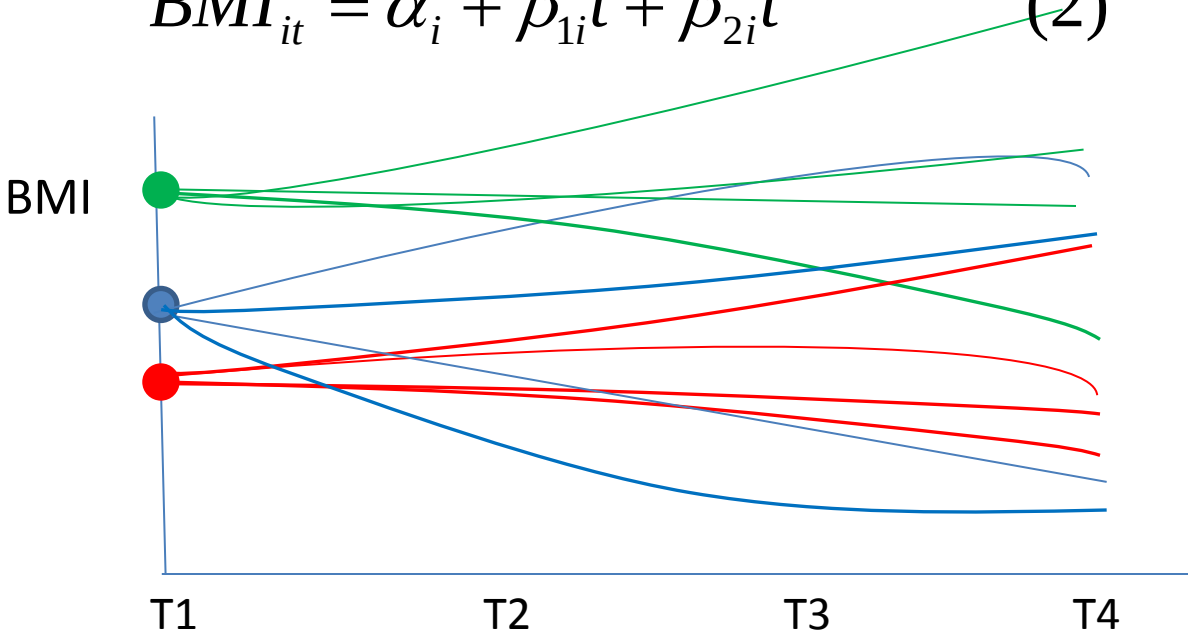
23

Why YES?

- The fitted line (Equation 1) provides a reference for each trajectory

$$BMI_t = \alpha + \beta_1 t + \beta_2 t^2 \qquad (1)$$

$$BMI_{it} = \alpha_i + \beta_{1i} t + \beta_{2i} t^2 \qquad (2)$$



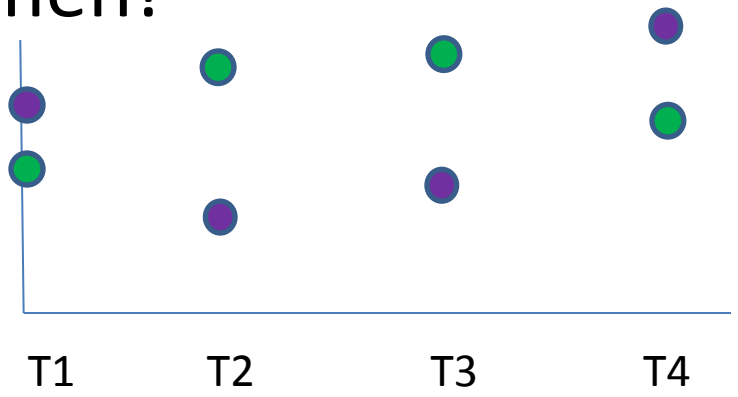
The question becomes:

- Can we select β_1 and β_2 such that the predicted BMI values of these individuals (Note: the ones surveyed at T1 only) at later times conform to some patterns?

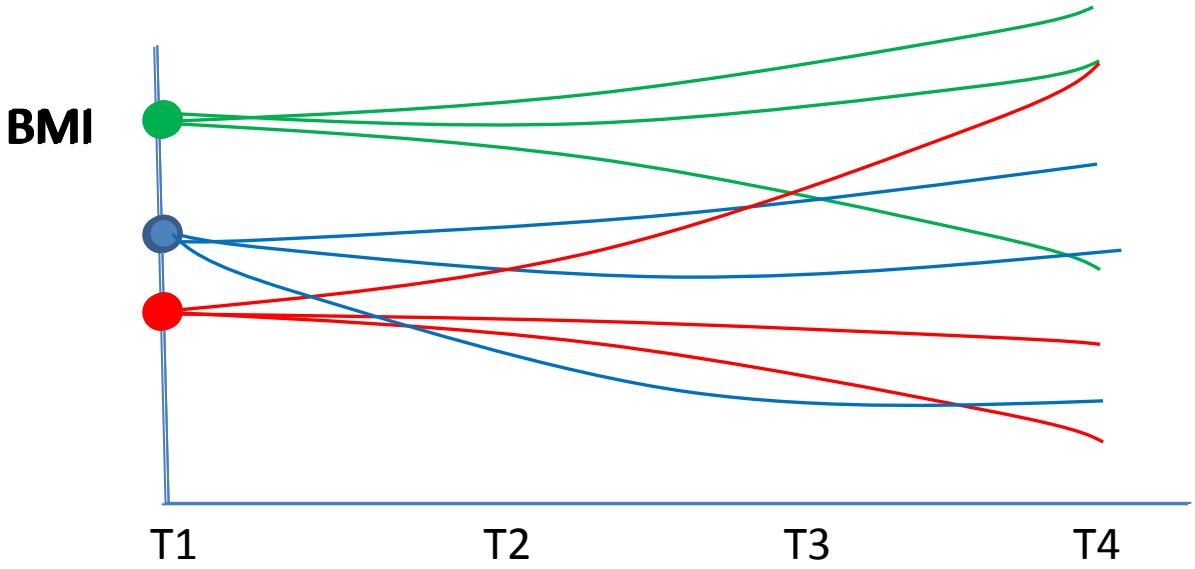
Step 4: Project BMI randomly

- For the two parameters β_1, and β_2, create the corresponding envelops $(\beta_1 - \Delta\beta_1, \beta_1 + \Delta\beta_1)$ and $(\beta_2 - \Delta\beta_2, \beta_2 + \Delta\beta_2)$
- From these two envelops we draw two random numbers to decide the trajectory
- Repeat it for all 845 women!

$$BMI_{it} = \alpha + \beta_{1i}t + \beta_{2i}t^2$$

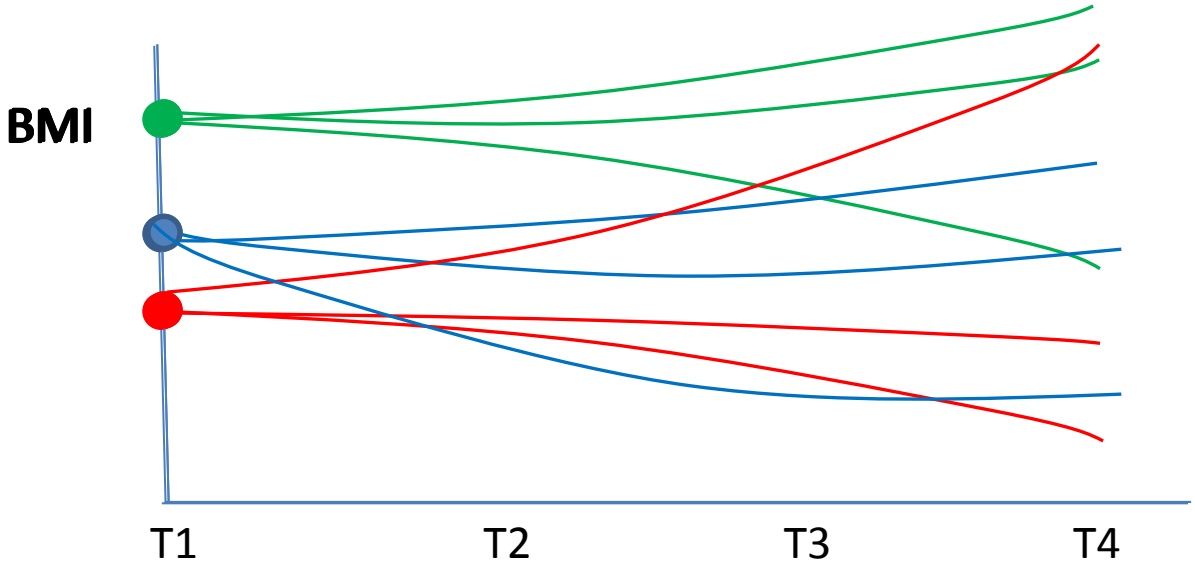


Step 5: Repeat Step 4 1000…0 times!



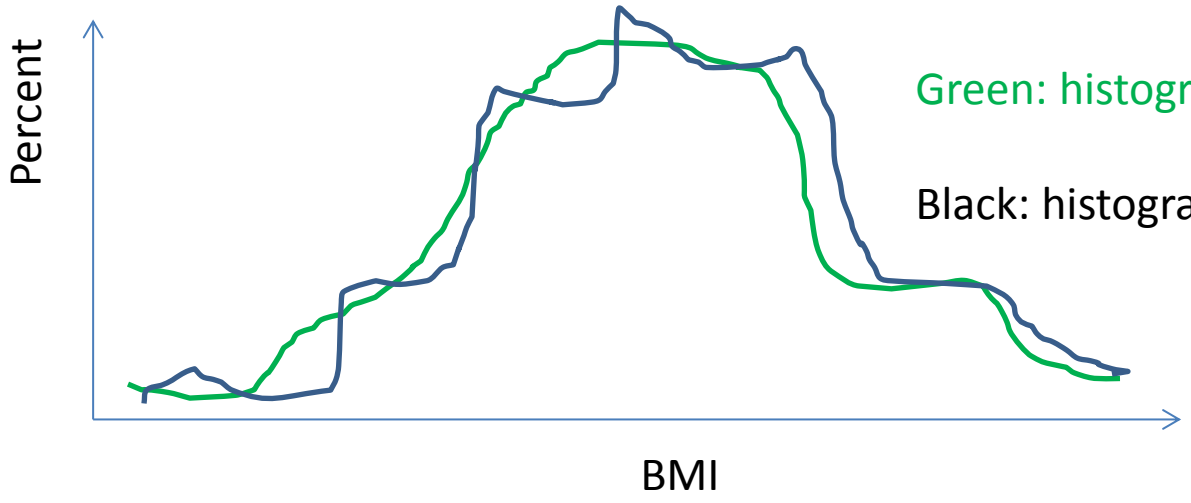
Out of these 1000…0 simulations

- Which one most resembles the real situation (i.e., if we had a chance to measure the BMI of all T1 individuals at T2, T3, and T4)? → Our job is to screen out all "bad" ones



Step 6: Histogram comparison

- At T2 (then T3, T4), we have real BMI measurements (but from different people) from DHS
- At T2 (then T3, T4), we have projected BMI values of the women who were surveyed at T1
- Question: Shall the two sets of BMI be similar in some aggregate statistical measures (e.g., histogram difference)?



Step 7 Parameter Regression

- Out of the 1000…0 randomly generated outcomes, the best outcome (the one that resembles the real situation) contains 985 sets of β_1 and β_2, which make the following regression model fit well:

$$\beta_1 = \mu_{\beta_1} + \gamma_{\beta_{1-1}} x_1 + \gamma_{\beta_{1-2}} x_2 + \zeta_{\beta_1} \quad (3)$$

$$\beta_2 = \mu_{\beta_2} + \gamma_{\beta_{2-1}} x_1 + \gamma_{\beta_{2-2}} x_2 + \zeta_{\beta_2} \quad (4)$$

where x_1 and x_2 are symbolically independent variables (could be more than two), the gammas (γ's) with subscripts are their coefficients, and zetas (ζs) are residuals

Parallel computing

- To generate these 1000…0 outcomes, calculate the histogram differences, and calculate regression fit (R^2), it takes at least one year to complete the computation in one PC
- Dr. Guangming He is performing parallel computing

Neighborhood impact

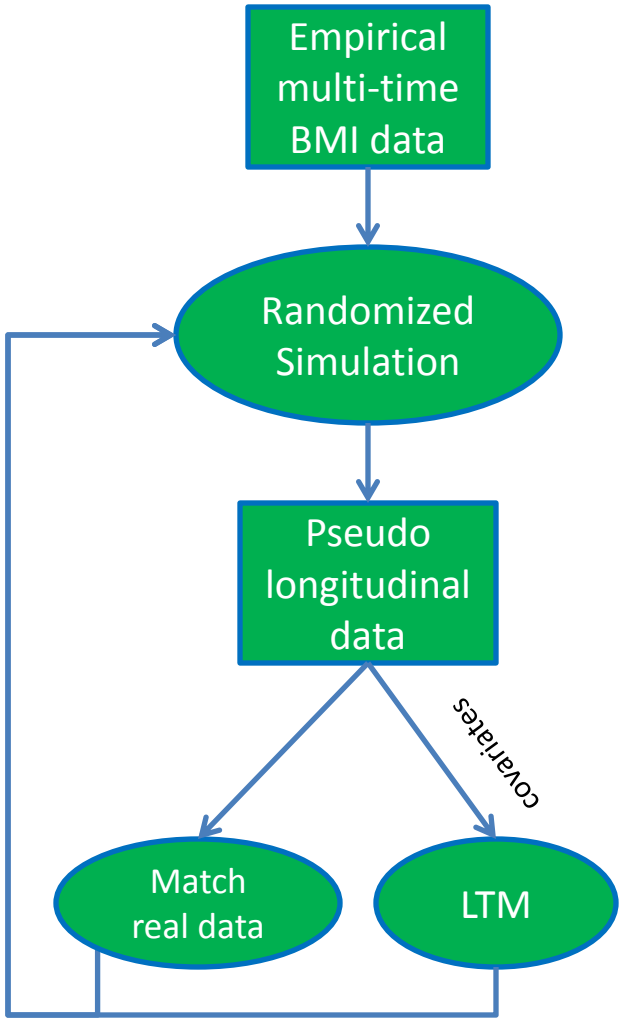
- In Step 7 regression, we put LULCC values obtained at 2.5 Km (same as Crook et al, 2016) → some data records share the same LULCC values
- Eigenvectors are used in Step 7 → some data records share the same eigenvector values
- A multilevel model is used to account for this within-cluster similarity.

Preliminary results

- Age has been found to be a positive predictor of BMI slope β_1
- Spouse at home has been found to be a positive predictor of BMI slope β_1
- More results are forthcoming

Methodology Significance

- Without true longitudinal data, we can still perform space-time analysis and find out temporal variability (without aggregating low-level to an upper level)
- Applications
  - Combine the results in an agent-based model (ABM)
  - Apply the method in Gallup data about climate change attitudes



Substance Significance

- Make full use of BMI data at HH level and explore what factors may affect BMI dynamics
- Understand the impact of age and other low-level variables in pseudo longitudinal data analysis
- Able to analyze the impacts of LULCC factors on BMI dynamics when age etc. in control

Questions?

- http://complexities.org/An/
- http://complexities.org

Acknowledgement

- The SDSU NASA project team: Doug Stow, John Weeks, Pete Coulter, Helena Taflin (Idea, data, and funding)
- Drs. Atsushi Nara, Andrew Zhang, and Ke Huang at SDSU (Python programming)
- Evan Casey at SDSU (Data preparation)
- Dr. Guangming He at Michigan State University (progamming and parallel computing)