Analytical approach for the IPEN project - Updates

Ester Cerin

C-PAN, Deakin University

IHP, The University of Hong Kong

IPEN
INTERNATIONAL PHYSICAL ACTIVITY AND THE ENVIRONMENT NETWORK

STEP 1 – Proposal formulation

- Aims and hypotheses (theoretically driven)
- Diagrammatic representation of hypotheses
- A list of selected variables (use code book!) and sites
- Proposed statistical methods and modeling approaches (these will vary depending on the aims of the paper [e.g., correlates or mediators? measurement or substantive paper?], outcome variables, and number of sites included in the analyses)
- Names/contact details of lead working group
- Names/contact details of analysts (if available)

A FEW UPDATES

Publication committee reviews the proposal and provides feedback on the modeling approach. Writing teams need to assess whether they have team members with appropriate expertise in the proposed analyses. The main analyst needs to be involved / provide input from the very start (STEP 1).

I cannot conduct or provide help with the analyses for all papers.

STEP 1 – Proposal formulation – example

UPDATES IN RED!!

- **A list of selected variables and sites**

SITES (currently available and including all variables of interest)

- Adelaide (Australia)
- Baltimore (USA)
- Ghent (Belgium)
- Seattle (USA)

SELECTED VARIABLES

- **Outcomes:** Total overall sitting (min/day; IPAQ-Long); Motorized transport time (min/wk; IPAQ-Long)
- **Explanatory variables:** NEWS-A: dwelling density; street connectivity; land use mix access; land use mix diversity (2 alternative forms); infrastructure for walking/cycling; traffic safety; crime safety; aesthetics; few cul-de-sacs; not many barriers to walking in neighborhood; parking difficult near shopping areas
- **Moderators:** gender and **site (ALWAYS)**
- **Covariates:** Age, gender, site, BMI, marital status, educational attainment, job status, **survey admin (in-person or self-report) (NEVER – SITE ARE FIXED EFFECTS; VARIABLES ARE PERFECTLY COLLINEAR)**
- **Sampling-related variables**: Participant ID, administrative unit, neighborhood-level SES and walkability (if explanatory variables do not capture these two aspects of the neighborhood environment)

STEP 1 – Proposal formulation – example

UPDATES IN RED!!

- Proposed statistical methods and modeling approaches
  - **Generalized additive mixed regression models (random intercepts)**
    - Multiple regression methods (multiple predictors)
    - Mixed models (multilevel models) to account for multiple levels of clustering (variation)
    - Generalized models to model non-normally distributed outcomes
    - Additive models to explore the shape of the relationships between explanatory variables and outcomes
  - **Modeling approach**
    - Identify redundant covariates; exclude them from models
    - Examine associations of single environmental attributes with outcomes (adjusted for study site and socio-demographic covariates) (**USE ALL ENTRY OR BACKWARD DELETION STRATEGY INSTEAD DUE TO POSSIBLE SUPPRESSION EFFECTS; START WITH THE MOST COMPLICATED MAIN EFFECT MODEL; THEN EXPLORE INTERACTIONS**)
      - Explore various variance (Gaussian, Gamma and inverse binomial) and link functions (logarithmic or identity)
      - Main effects and two-way and three-way interaction effects (gender and site)
    - Examine independent association of multiple environmental attributes with outcomes **(SEE COMMENT ABOVE)**
      - Use best fitting variance and link functions from single-predictor models
      - Include significant main and interaction effects (from previous models)
    - Construct composite indices of neighborhood walkability and PA friendliness based on multiple-predictor models (**WHEN APPROPRIATE; LESS APPROPRIATE WHEN DEALING WITH MULTIPLE MEASURES OF THE SAME CONSTRUCT WITH DIVERSE PATTERNS OF SIGNIFICANT PREDICTORS**)
      - Use best fitting variance and link functions from single-predictor models
      - Main effects and two-way and three-way interaction effects (gender and site)

STEP 1 – Proposal formulation – example

UPDATES IN RED!!

- Proposed statistical methods and modeling approaches
  - **Modeling approach**
    - Dealing with an excessive number of zeros
      - Model of participation vs non-participation in specific type of activity (binomial variance and logit link function)
      - Model of non-zero values (e.g., non-zero frequency and minutes of walking for recreation)
      - Figuring out how to model these simultaneously ... Will keep you updated
    - Example (Paper 3) – Perceived environmental attributes and walking for recreation
      - Inspection of residual plots indicating problems with distributions
      - Zero-inflated GAM indicating need for zero-inflated models (Vuong's test significant)
        - Zero-inflated GAMMs not available
        - Separate models for non-zero outcome values and zero vs. non-zero outcome values

STEPS 1 & 4 – Proposal formulation and data analyses – example

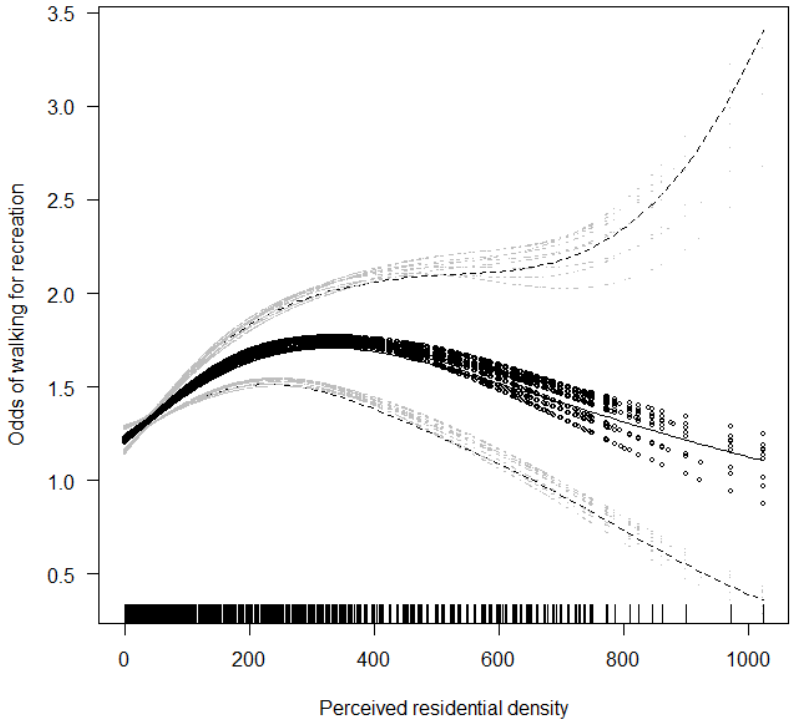
UPDATES ...

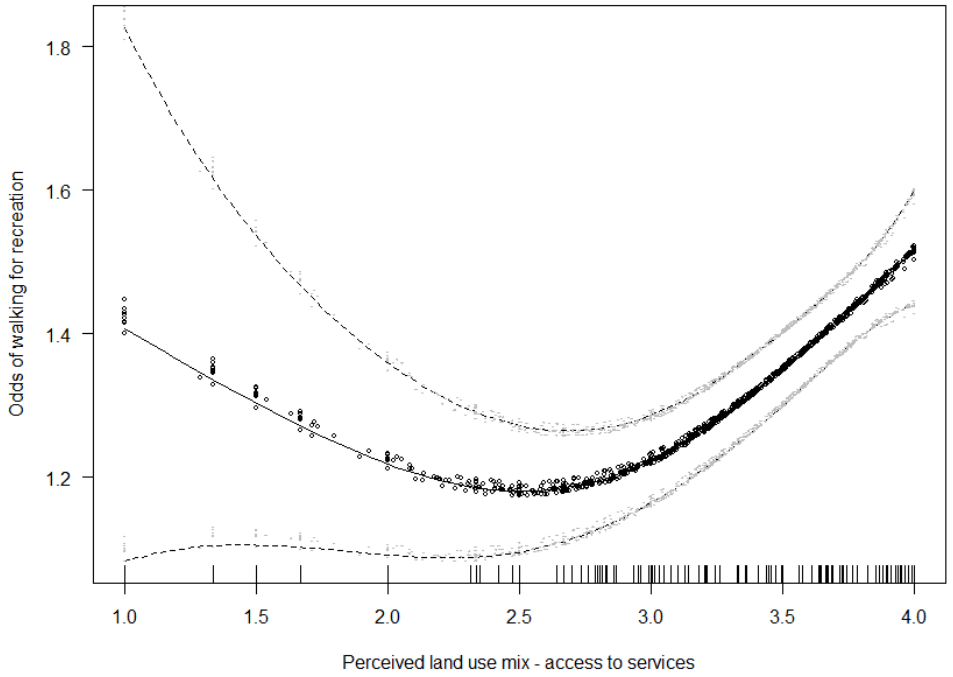
Linear and curvilinear associations of perceived environmental attributes with recreational walking (main effects)

Perceived environmental attribute	Odds of walking for recreation[a] (N=13745)			Non-zero frequency of walking for recreation[b] (N=7838)			Non-zero minutes of walking for recreation[b] (N=7838)		
	OR	95% CI	p	exp(b)	exp(95% CI)	p	exp(b)	exp(95% CI)	p
Residential density Curvilinear component	1.08 F(2.57)=7.94	0.88, 1.32	.482 <.001	1.001 -	0.999, 1.003 -	.138 -	1.001 -	1.000, 1.001 -	.005 -
Land use mix – access Curvilinear component	1.02 F(2.49)=8.95	0.90, 1.17	.726 <.001	1.02 -	0.99, 1.05 -	.062 -	1.07 -	1.02, 1.11 -	.003
Connectivity	1.01	0.96, 1.07	.745	1.02	1.00, 1.05	.025	1.02	0.98, 1.05	.298
Infrastructure and safety	1.01	0.94, 1.09	.826	0.97	0.94, 1.00	.053	0.98	0.92, 1.04	.519
Aesthetics Curvilinear component	1.26 -	1.18, 1.35 -	<.001 -	1.05 -	1.03, 1.08 -	<.001 -	1.02 F(2.28)=6.56	0.95, 1.10	.569 <.001
Safety from traffic	1.05	0.99, 1.11	.128	0.99	0.96, 1.01	.313	0.97	0.93, 1.01	.101
Safety from crime	1.07	1.00, 1.14	.036	1.00	0.97, 1.02	.737	0.98	0.94, 1.02	.334
Few cul-de-sacs	0.94	0.90, 0.98	.024	0.99	0.98, 1.01	.274	0.98	0.95, 1.00	.059
No major barriers	1.00	0.95, 1.05	.886	1.00	0.98, 1.02	.898	1.00	0.97, 1.04	.795
Proximity to parks	1.07	1.03, 1.11	.031	1.01	0.99, 1.02	.213	1.01	0.99, 1.04	.344

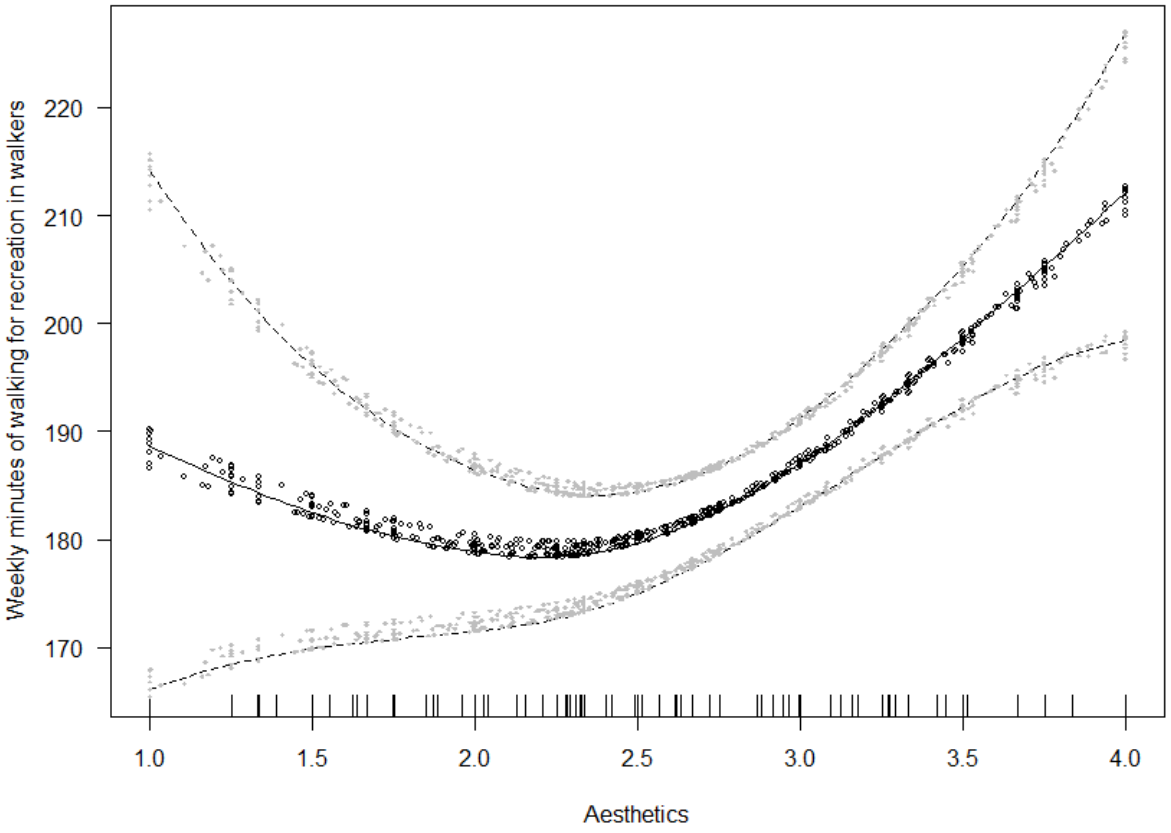
UPDATES ...

Non-linear relationship between environmental attributes and the odds of walking for recreation





Non-linear relationship between perceived aesthetics and non-zero weekly minutes of walking for recreation



UPDATES ...

STEP 2 – Data preparation

- Coordinating center:
  - cleans each dataset
  - sends merged dataset with codebook

- **If needed, writing team creates new variables to be used in the proposed models (including recoding; creation of new composite variables)**

- Create a codebook for new variables

- **Run descriptive statistics on the merged dataset to identify % of missing values and obtain main descriptive statistics (means, SD, medians, %, etc.)**

- Verify the validity of the data for newly-created variables (out-of-range values, etc.)

NO UPDATES …

I don't think I need to provide assistance with STEP 2 (unless I'm the only or primary analyst conducting the analyses)

STEP 3 – Multiple imputation models

UPDATES ...

- Create at least 10 imputed datasets if 5-10% of cases have missing values on at least one variable (5-10% listwise missing data for the variables being examined in the paper); if >10% missing data: create as many imputed datasets as the percentage of missing data (20% missing data = 20 imputed datasets)

- **Why?**
  - Most data are missing at random (MAR): the probability that an observation is missing commonly depends on information that is available in the dataset, i.e., the reason for missingness is based on other observed participant's characteristics. In this case, participants with complete data are a biased sub-sample of all participants. A complete case analysis would usually produce biased results. In addition, there is loss of power due to analyzing a smaller dataset.
  - All simple techniques of handling missing data (complete case analysis, the indicator method and overall mean imputation) give biased results.
  - Multiple imputations are a more appropriate alternative because they produce good estimates of variability in the dataset and yield unbiased estimates if missing data are MAR.

Questions?

- SOFTWARE
- Modeling
- Interpretation of findings
- Presentation of findings

IPEN
INTERNATIONAL PHYSICAL ACTIVITY AND THE ENVIRONMENT NETWORK