McMaster University

MacDATA FELLOWSHIP PROPOSAL

A multivariate multilevel analysis for binary outcomes: correlates of sanitary improvements in the slums of Nairobi

Student:
Steve Cygu
Student number:
400164479

Supervisory committee members:
Prof. Jonathan Dushoff
Prof. Ben Bolker
Prof. James Reilly
Dr. Seow Hsien
Prof. Paul McNicholas

March 4, 2021

1 Introduction

Many residents of urban areas face joint obstacles to basic service access, including social-economic, institutional, spatial and political barriers, and these are even more prevalent and severe in slum settlements (Pierce, 2017). In Kenya, the Nairobi Urban Health and Demographic Surveillance System (NUHDSS) has collected data since 2003 on household-level access to WaSH services in two large slum areas (Beguy et al., 2015). The NUHDSS data provides a useful starting point for understanding the factors associated with trends in access to WaSH services in slum areas.

In studying WaSH services, we have three variables (improved water, toilet facilities and garbage collection) characterizing our main outcome. The most common approach to analyzing multiple outcomes is to analyze each outcome independently, in a univariate framework. This approach ignores the most likely unmeasured variations and correlation among the outcomes and the multivariate structure of the data. In addition, in situations where multiple binary outcomes are simultaneously observed in longitudinal (for example, at household level) data presents some modelling challenges, since the models should reflect that:

1. The outcomes for each household are likely to be correlated.
2. The multiple outcomes reflect many of the same underlying processes.
3. Outcome-specific effect sizes may be of interest.

A naive attempt to overcome 1 and 2 would be combining outcome indicators into scores (representing an aggregated assessment of outcomes for each household), and then regressing the scores against the covariates. One major limitation of these approaches is that, by focussing on the aggregate, they fail to address 3, i.e., they lack the ability to identify which particular outcome is associated with the effect of interest. Structural equation models (SEMs) and explanatory SEMs are alternative approaches to addressing all the three that introduce latent variables to jointly model multiple outcomes (Das et al.; Fang et al., 2018; Ivanova et al., 2016; Miro-Quesada et al., 2004; LaLonde et al., 2019). However, latent variables approach generally involve strong statistical assumptions which may have a strong influence on the findings. LaLonde et al. (2019) discusses some of the major drawbacks of using latent variable approaches to jointly model multiple outcome models.

Individual modeling of each outcome – with separate mixed logistic linear models being fitted for each outcome, would attempt to solve 3 at the expense of 1 and 2. Specifically, separate models will ignore the fact that the outcomes observed from the same subject (household) are likely to be correlated, since they are subject to shared influences that are distinctive to that particular household. In other words, fitting separate models to several related outcomes cannot address the question whether there is a global relation between the outcomes (LaLonde et al., 2019). Ignoring such correlations may lead to poor estimates (Das et al.; Fang et al., 2018; Ivanova et al., 2016; Miro-Quesada et al., 2004; LaLonde et al., 2019).

Joint modeling of multiple outcomes is sometimes preferable to separate models especially in longitudinal studies (Das et al.; Fang et al., 2018; LaLonde et al., 2019). There are a number of advantages to this approach. First, in spite of its simultaneous formulation, by correctly nesting

the effect of interest (for example, a particular covariate) within the multiple outcomes, both outcome-specific and global effects can be estimated (Das et al.; LaLonde et al., 2019). Second, association between outcomes can be captured in terms of correlation between household-level random effects (Ivanova et al., 2016). Third, joint modeling may increase statistical power (Das et al.).

2 Research goals

In this project we will adopt a joint modeling approach to analyse binary outcomes. Our approach will take into account the longitudinal nature of the data to simultaneously model all the three WaSH outcome variables (improved water, toilet facilities and garbage collection). Specifically, this project aims to investigate the contribution of demographic and economic factors to improved water, toilet facilities and garbage collection among the Nairobi urban poor using NUHDSS data.

3 Interdisciplinary nature

By formulation, joint modeling of WaSH indicators inherently requires interdisciplinary approach because of its intersection with several paradigms – public health and global health, statistics and data science. In particular, the proposed approach is informed by the domain expertise, for example, on how the WaSH indicators could potentially correlate within households. Consequently, the perspective of this project is to incorporate this information with an aim of improving on various data retrieval, handling, visualization techniques and statistical methods while examining the contribution of demographic and economic factors to improved WaSH indicators.

4 Methodology

4.1 Data acquisition

This project with use existing data from a longitudinal NUHDSS covering two major urban slums in Nairobi, Kenya. The baseline survey that defined the initial population for the NUHDSS was carried out from July–August 2002. Subsequently, demographic, socio-economic, household characteristics and livelihood sources data were yearly collected until the end of 2015. Beguy et al. (2015) gives a complete description of the NUHDSS study design and setting.

Outcomes

We will focus on three WaSH variables classified as improved or unimproved – this follows WHO guidelines (Yu et al., 2016).

Table 1: Multiple outcomes categorization.

	Improved	**Unimproved**
Drinking water source	Piped water into dwelling, plot or yard Public tap/standpipe Tube well / borehole Protected dug well with hand pump Protected spring Rainwater collection from the roof	Unprotected dug well Unprotected spring Small water vendor (cart with small tank or drum) Bottled water Tanker truck Rainwater collection from surface run off. Protected dug well with bucket
Toilet facility type	Flush / pour flush to piped sewer system or septic tank or pit latrine VIP latrine Pit latrine with slab Composting toilet	Flush / pour flush to elsewhere e.g. to open drain Pit latrine without slab (slab with holes) /open pit Bucket Hanging toilet / hanging latrine No facilities or bush or field
Garbage disposal method	Garbage dump Private pits Public pits Proper garbage disposal services Other organized groups such as the national youth service	In the river On the road, railway line/station In drainage/sewage/trench Vacant/abandoned house/plot/field No designated place/all over Street boys/urchins Burning Other

Covariates

Below are some of the covariates which will be included in our model:

- **Demographics**
  - Age
  - Gender
  - Slum area (Korogocho and Viwandani)
  - Number of household members (in this + other structure)
  - The interview year
- **Dwelling index**

  Dwelling index will be derived from the first principal component score of the PCA based on dummy indicator variables generated from the following household amenities questions, i.e., main material of the
  - floor
  - roof
  - wall
  - fuel
  - lightning
  - ownership
- **Assets ownership index**

Assets information on whether the household owned any (yes/no), either at the household or elsewhere, will be used to perform logistic PCA to and then first PC scores used as a proxy for assets ownership index.

- **Household income**

  Estimated total income for the household in the last 30 days.

- **Total expenditure**

  Sum of amount (KES) spent on the following, in the last 7 days, constitutes household total expenditure.

- **Household food consumption**

  How do you describe the food eaten by your household in the last 30 days?

- **Shocks index**

  Whether the household experienced any of the following shocks/problems in the last year (yes/no)?

- **Household self rating**

  On a scale of 1 (poorest) - 10 (richest), how does the household compare to others in the community.

4.2 Analysis plan

We will fit a generalized mixed model (using both classical frequentist mixed-model and Bayesian frameworks), i.e., joint hierarchical logistic regression model with shared random effects that accounts for the correlation due to households and years. Specifically, the shared random intercept will account for the correlation between the three outcomes of the same household in a particular year, and captures the unobserved factors specific to each household (in a particular year) that may influence the three services.

In order to explore and understand the two fitting approaches, we will perform simulation-based validation. The main idea is to validate the proposed model and redefine the it, if necessary using simulated data.

5 Significance

While WaSH services constitute some of the most basic requirements for human health and dignity throughout the world, this dignity is missing in most of the slum areas of Nairobi. Even though there have been a number of interventions intended to upgrade Nairobi's slum areas, focusing on issues of infrastructure development, especially in the Korogocho and Viwandani slums, the communities remain exposed to overcrowding and poverty, alcohol and substance abuse, domestic violence and crime. In addition to widespread poverty, residents of Nairobi's Viwandani and Korogocho slum areas are also faced by the near-absence of most of the basic

services they need to live healthy lives. It is also important to note that Kenya has experienced unprecedented urban growth, which is expected to lead to the country's urban population reaching about 31.7 million (56%) by 2027. This rapid urbanization has left Kenyan cities with a huge unmet demand for critical infrastructure and basic services, adversely affecting the quality of life for urban residents, with nearly two-thirds of urban residents having no access to improved sanitation (Chikozho et al., 2019). The result from this study will provide an overview to the extent to which residents of Nairobi's slum areas of Viwandani and Korogocho have been able to access and the status of WaSH services. In addition, the result is intended to inform the agenda of policymakers and public health practitioners who grapple continuously with the challenges faced in accessing WaSH services in Nairobi's low-income residential areas. It will also directly contributes to the growing body of knowledge on access to improved WaSH services in the context of slum areas in low and middle-income countries.

6 Distinction between MacDATA proposal and PhD project

Many techniques from statistical methods (SM) and machine learning (ML) may, in principle, be used for both perspectives. However, SM has a well established focus on inference by building probabilistic models which allows us to determine a quantitative measure of confidence about the clarity of the true effect. Simulation-based validation approaches can be used in conjunction with SM to explicitly verify assumptions and redefine the specified model, if necessary. On the other hand, ML uses general-purpose algorithms to find patterns that best predict the outcome and makes minimal assumptions about the data-generating process; and may be more effective in a number of situations (Bzdok et al., 2018). However, despite convincing predictive performance and flexibility, the lack of explicit models in most ML methods can make ML results difficult to directly link to prior public health knowledge.

Owing to the differing strengths of the two approaches, we propose two computational approaches for investigating particular public health. In particular:

- Aims and scope of PhD thesis
  - Using machine learning methods to build, validate and compare prognostic ML models to predict survival in patients with cancer using both clinical and patient reported outcomes, and incorporating changes in the measured covariates over time using unique databases available in Ontario, Canada
- Aims and scope of MacDATA project
  - A multivariate multilevel analysis for binary outcomes: correlates of sanitary improvements in the slums of Nairobi

The first case focuses on ML-based cross-validation to evaluate predictive models for maximized predictive performance on unobserved outcomes. In the second case we propose statistical modeling together with simulation-based validation approaches to majorly focus on causal inference of model parameters.

7 MacDATA Graduate Fellowship supervisors

1. **Dr. Jonathan Dushoff**

   Professor, Department of Department of Biology, McMaster University

   Faculty Affiliate: Department of Computational Science and Engineering, MacMaster University

   Life Sciences Building, LSB 332, McMaster University, 1280 Main Street West

   Email: dushoff@mcmaster.ca

   Office Phone: (905) 525-9140 ext. 26313

   Website: https://mac-theobio.github.io/dushoff.html

   MacDATA member: yes

2. ...

References

D. Beguy, P. Elung'ata, B. Mberu, C. Oduor, M. Wamukoya, B. Nganyi, and A. Ezeh. Health & demographic surveillance system profile: the nairobi urban health and demographic surveillance system (nuhdss). *International journal of epidemiology*, 44(2):462–471, 2015.

D. Bzdok, N. Altman, and M. Krzywinski. Points of significance: statistics versus machine learning, 2018.

C. Chikozho, D. T. Kadengye, M. Wamukoya, and B. O. Orindi. Leaving no one behind? analysis of trends in access to water and sanitation services in the slum areas of nairobi, 2003–2015. *Journal of Water, Sanitation and Hygiene for Development*, 9(3):549–558, 2019.

A. Das, W. K. Poole, and H. S. Bada. Simultaneous modeling of multiple binary outcomes: A repeated measures approach.

D. Fang, R. Sun, and J. R. Wilson. Joint modeling of correlated binary outcomes: The case of contraceptive use and hiv knowledge in bangladesh. *PloS one*, 13(1):e0190917, 2018.

A. Ivanova, G. Molenberghs, and G. Verbeke. Mixed models approaches for joint modeling of different types of responses. *Journal of Biopharmaceutical Statistics*, 26(4):601–618, 2016.

A. LaLonde, T. Love, S. W. Thurston, and P. W. Davidson. Discovering structure in multiple outcomes models for tests of childhood neurodevelopment. *Biometrics*, November 2019. ISSN 0006-341X. doi: 10.1111/biom.13174. URL https://doi.org/10.1111/biom.13174.

G. Miro-Quesada, E. Del Castillo, and J. J. Peterson. A bayesian approach for multiple response surface optimization in the presence of noise variables. *Journal of applied statistics*, 31(3): 251–270, 2004.

G. Pierce. Why is basic service access worse in slums? a synthesis of obstacles. *Development in Practice*, 27(3):288–300, 2017.

W. Yu, N. A. Wardrop, R. E. S. Bain, Y. Lin, C. Zhang, and J. A. Wright. A global perspective on drinking-water and sanitation classification: An evaluation of census content. *PLOS ONE*, 11(3):1–17, 03 2016. doi: 10.1371/journal.pone.0151645. URL https://doi.org/10.1371/journal.pone.0151645.