MacDATA FELLOWSHIP PROPOSAL

A multivariate longitudinal analysis with 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 15, 2021

Introduction

Many residents of urban areas face multiple 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 water, sanitation and hygiene (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.

Our main study outcome is measured by three variables (improved water, toilet facilities and garbage collection). This is a challenging data set for a number of reasons. We expect outcome variables to be correlated, beyond factors explained by our predictors, due to causal interactions and unmeansured covariates). We also expect temporal correlations between measurements from each household. Dealing with these considerations is complicated by the fact that the outcome variables are binary, and we thus cannot use standard techniques based on outcome normality. We would like a modeling approach that can:

1. Control for correlations within househoulds
2. Address correlation between outcomes
3. Estimate outcome-specific effects

One way to address 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. This method fails to address 3: it shows only how predictors connect to the aggregate outcome, not with particular outcomes. Univariate-response models – with separate mixed logistic linear models being fitted for each outcome, for example – address 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 (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.). But as noted above, correctly formulating and fitting these model is particularly challenging in our case of binary outcomes.

Research goals

We will develop, validate and apply a joint modeling approach to analyse binary outcomes. Our approach will take into account the longitudinal nature of the data to model all three WaSH outcome variables (improved water, toilet facilities and garbage collection). The analysis will be based on a generalized linear mixed model approach. We will compare univariate- and multivariate-outcome models. By investigating the contribution of demographic and economic factors to WaSH access, we hope to identify overlooked factors that will be helpful to government and other development workers seeking to extend and equalize access.

Interdisciplinary nature

Multivariate modeling of WaSH indicators is an inherently interdisciplinary question because of its intersection with several paradigms – public health, statistics and data science. The proposed approach will be informed by public-health expertise – for example, on what factors may lead to correlations in the outcome variables beyond those explained by predictors. The candidate, supervisors, and committee will work with subject-matter experts based in Kenya both to refine the analysis approach and to interpret the results and their potential implications for improving development and for guiding future research questions. The statistical questions of validating and correctly applying these applying are at the heart of the project. Data science methods will also be important. We will carefully clean and curate the NUHDSS data, and will aim to have a complete analysis pipeline that will make it easy to replicate our analysis with updated data, or with different assumptions, should that become necessary.

Methodology

Data acquisition

We will 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 collected yearly 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: Drinking water source; Toilet facility type; and Garbage disposal method. Each of these is classified as improved or unimproved – this follows WHO guidelines (Yu et al., 2016).

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
- **Socio-economic**
  - Dwelling index
  - Assets ownership index
  - Household income
  - Total expenditure
  - Household food consumption
  - Shocks index
  - Household self rating

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 correlated random intercepts will account for the correlation between the three outcomes of the same household in a particular year, and capture 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 use simulations that match our assumptions to validate and refine the proposed models before applying them to the real data set.

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 results 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. We will work with Kenya-based subject-matter experts to frame the study; interpret the results; and plan future research based on these resutls. We are hopeful that this research can be used 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.

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`.

Distinction between MacDATA proposal and PhD project

Many techniques from statistical methods (SM) and machine learning (ML) overlap. 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. Despite convincing predictive performance and flexibility, however, the lack of explicit models in most ML methods can make ML results difficult to directly link to prior public health knowledge.

The candidate's PhD thesis and the proposed project share the idea of validated learning about public-health questions, but they emply different perspectives, and diferent computational approaches, which are applied to different questions:

- 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 focus on causal inference of model parameters.

Steve Bicko Cygu

4 Beaucourt Pl, Hamilton, ON L8S 2P7

☎ +1 289 55 66 241 ✉ cygus@mcmaster.ca

Research Goals

I am an applied statistician with strong background in mathematics, computing and data science. My focus is on application of machine learning tools in trying to understand public health challenges. Specifically, using machine learning methods to predict survival of cancer patients.

Education

2018 – Date	**PhD in Computational Science and Engineering, McMaster University, Canada.**
2016 – 2017	MSc in Mathematics, Stellenbosch University, South Africa.
2014 – 2015	MSc in Mathematical Sciences, University of Cape Town, South Africa.
2009 – 2013	BSc in Applied Statistics with Computing, Maasai Mara University, Kenya.
2004 – 2007	Kenya Certificate of Secondary Education, Waondo Secondary School, Kenya.

Research Experience

2018 – Date **Doctoral Candidate, McMaster University, Canada.**

- Using Machine Learning Methods to Predict the Risk of Death of Gastrointestinal Cancer Patients
- Modeling approaches for multivariate binary response. A case study to investigate the contribution of demographic, social and economic factors to improved water, sanitation and hygiene among the urban poor

2016 – 2017 **Graduate Student, Stellenbosch University, Canada.**

- Immune biomarker reference range estimation for healthy paediatric patients in South Africa
- Determining paediatric Immune Biomarker Reference Ranges using a model-based, age-continuous estimation method

Publications

arXiv preprint	**Cygu, Steve**, and Benjamin M. Bolker. *”pcoxtime: Penalized Cox Proportional Hazard Model for Time-dependent Covariates.” arXiv preprint arXiv:2102.02297 (2021)*
Manuscript	**Steve Cygu**, Helen Payne, Denise Lawrie, Debbie Glencross, Martin Nieuwoudt. *Determining paediatric Immune Biomarker Reference Ranges using a model-based, age-continuous estimation method*

Teaching Experience

2018 – Date **McMaster University, Canada.**

Teaching assistant for

- Introduction to mathematical and scientific programming
- Population ecology

Professional Experience

May – Aug., 2018 **African Population and Health Research Center (APHRC), Kenya.**

Data Analyst

- Developed data visualization tools in R
- Developed tools to monitor and clean in coming data in realtime
- Insightful analysis of data collected
- Questionnaire scripting and development of online based data collection tools
- Data cleaning and management

Mar, 2017 – May., 2018 **Dalberg Research, Kenya.**

Data Processing Manager

- Developed automated data quality check scripts
- Coordination and management of data management team
- Provided technical assistance to clients
- Data analysis and management
- Developed data management SOPs

2016 – 2017 **South African Centre of Excellence in Epidemiology Modelling and Analysis, South Africa.**

Research Student

- Developed statistical and mathematical models for the analysis of epidemiological data
- Developed R software package for the estimation of reference ranges

Jan., 2014 – Sept., 2014 **Infotrak Research and Consulting, Kenya.**

Data Processing Manager

- Designed survey tools
- Supervised data collection and entry process
- Managed databases and provided realtime update on survey progress
- Conducted data cleaning, processing and analysis
- provided technical assistance to business development team

Academic Administrative Experience

2012 – 2013	Student Representative (Academic Director), Maasai Mara University.
2011	Treasurer Information Club, Maasai Mara University.
2010	Chairman Mathematics club, Maasai Mara University.

Awards

2016	South African Centre of Excellence Epidemiological Modeling and Analysis Scholarship
2014	frican Institute for Mathematical Sciences (Next Einstein Initiative) Scholarship

Conferences Attended

2019	Compute Ontario HPC Summer School, Hamilton, Canada
2016	Bayesian Analysis of Longitudinal Studies, Stellenbosch, South Africa
	Quantitative Bias Analysis with Epidemiological Data, Stellenbosch, South Africa
	Mathematical Modeling for Infectious Diseases, Cape Town, South Africa
	Clinic on Meaningful Modeling of Epidemiological Data (MMED)

Memberships

2018 – Date	Kenya Statistical Society
2016 – Date	South African Statistical Association (SASA)

Languages

Luo	Mother tongue
Swahili	Fluent
English	Fluent

Referees

PhD supervisor	**Prof. Jonathan Dushoff**, Department of Biology, McMaster University Email: dushoff@mcmaster.ca
MacData co-supervisor	**Prof. Harry Shannon**, Faculty of Health Sciences, McMaster University Email: shannonh@mcmaster.ca

MacDATA Graduate Fellowship supervisors

1. **Dr. Jonathan Dushoff**

   Professor, Department of Biology, McMaster University; Affiliate, Department of Computational Science and Engineering; Member, MacDATA

   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`

2. **Dr. Harry Shannon**

   Professor Emeritus, Health Research Methods, Evidence and Impact, Faculty of Health Sciences

   Faculty Affiliate: Faculty of Health Sciences, McMaster University

   Email: `shannonh@mcmaster.ca`

   Office Phone: (905) 525-9140 ext. 22147

   Website: `https://globalhealth.mcmaster.ca/people/harry-shannon`

Student number: 400164479