Working toward effective anonymization for surveillance data:
Innovation at South Africa's Agincourt Health and Demographic Surveillance Site

- Lori M. Hunter, Wayne Twine and Catherine Talbot

Extended Abstract

Research on population-environment connections is often hampered by lack of the necessary geographic coding required to link people to place.

To facilitate research on socio-ecological intersections in low-income settings, this project offers innovations in data integration through a focus on one of 50+ "Health and Demographic Surveillance Systems" providing unparalleled social science data in hard-to-reach world regions. This paper reports on 3 goals.

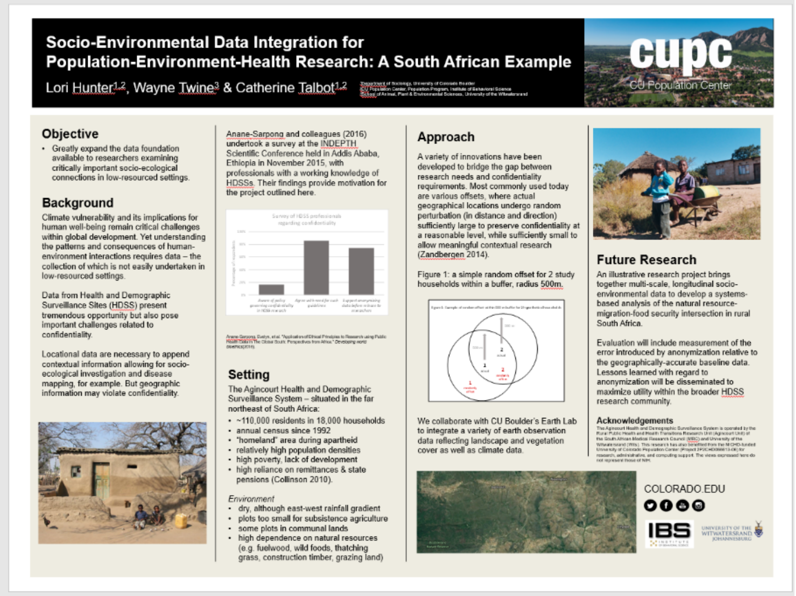
First, we **anonymize** household-scale data in our rural South African study setting, the Agincourt Health and Demographic Surveillance site. Integration of local environmental information with household data requires household locational information – but such information could breach confidentiality. Anonymization involves small random perturbations of locational data to protect household confidentiality.

Using clusters of household defined by differing rules, we then **integrate** earth observational information to measure local environmental conditions and temporal change.

Third, the **empirical effects** of different approaches to clustering will be explored to determine those which best balance research and confidentiality requirements. We will examine the intersections between migration, natural resource availability and food security using natural resource measures based on different clustering approaches.

In the long-run, the goal is to develop effective and efficient options for the provision of geocodes to researchers for socioecological data integration within Agincourt and potentially within other research settings.

We have already done a poster presentation of an early version of this project at CU Population Center's mini-conference on Africa Population-Health-Environment in April 2018. Since then, we have generated different clustering approaches and are implementing those now. The clusters will then be linked with earth observational data, measuring natural resources, which we already have in-hand. Results across models of migration, natural resources and household well-being (namely, food security) will be contrasted and provide the focus of this presentation.



Socio-Environmental Data Integration for Population-Environment-Health Research: A South African Example

Lori Hunter[1,2], Wayne Twine[3] & Catherine Talbot[1,2]

[1]Department of Sociology, University of Colorado Boulder
[2]CU Population Center, Population Program, Institute of Behavioral Science
[3]School of Animal, Plant & Environmental Sciences, University of the Witwatersrand

Objective

- Greatly expand the data foundation available to researchers examining critically important socio-ecological connections in low-resourced settings.

Background

Climate vulnerability and its implications for human well-being remain critical challenges within global development. Yet understanding the patterns and consequences of human-environment interactions requires data – the collection of which is not easily undertaken in low-resourced settings.

Data from Health and Demographic Surveillance Sites (HDSS) present tremendous opportunity but also pose important challenges related to confidentiality.

Locational data are necessary to append contextual information allowing for socio-ecological investigation and disease mapping, for example. But geographic information may violate confidentiality.

Anane-Sarpong and colleagues (2016) undertook a survey at the INDEPTH Scientific Conference held in Addis Ababa, Ethiopia in November 2015, with professionals with a working knowledge of HDSSs. Their findings provide motivation for the project outlined here.

Survey of HDSS professionals regarding confidentiality
Percentage of respondents
100%
80%
60%
40%
20%
0%
Aware of policy governing confidentiality in HDSS research
Agree with need for such guidelines
Support anonymizing data before release to researchers

Anane-Sarpong, Evelyn, et al. "Application of Ethical Principles to Research using Public Health Data in The Global South: Perspectives from Africa." *Developing world bioethics* (2016).

Setting

The Agincourt Health and Demographic Surveillance System – situated in the far northeast of South Africa:

- ~110,000 residents in 18,000 households
- annual census since 1992
- "homeland" area during apartheid
- relatively high population densities
- high poverty, lack of development
- high reliance on remittances & state pensions (Collinson 2010).

Environment

- dry, although east-west rainfall gradient
- plots too small for subsistence agriculture
- some plots in communal lands
- high dependence on natural resources (e.g. fuelwood, wild foods, thatching grass, construction timber, grazing land)

Approach

A variety of innovations have been developed to bridge the gap between research needs and confidentiality requirements. Most commonly used today are various offsets, where actual geographical locations undergo random perturbation (in distance and direction) sufficiently large to preserve confidentiality at a reasonable level, while sufficiently small to allow meaningful contextual research (Zandbergen 2014).

Figure 1: a simple random offset for 2 study households within a buffer, radius 500m.

Figure 1: Example of random offset within 500 m buffer for 2 hypothetical households
500 m
500 m
2
actual
1
actual
2
randomly offset
1
randomly offset

We collaborate with CU Boulder's Earth Lab to integrate a variety of earth observation data reflecting landscape and vegetation cover as well as climate data.

Future Research

An illustrative research project brings together multi-scale, longitudinal socio-environmental data to develop a systems-based analysis of the natural resource-migration-food security intersection in rural South Africa.

Evaluation will include measurement of the error introduced by anonymization relative to the geographically-accurate baseline data. Lessons learned with regard to anonymization will be disseminated to maximize utility within the broader HDSS research community.

Acknowledgements

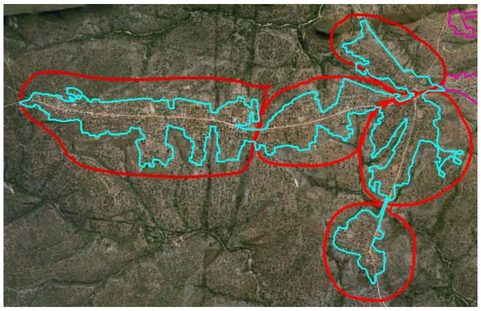
The Agincourt Health and Demographic Surveillance System is operated by the Rural Public Health and Health Transitions Research Unit (Agincourt Unit) of the South African Medical Research Council (MRC) and University of the Witwatersrand (Wits). This research has also benefited from the NICHD-funded University of Colorado Population Center (Project 2P2CHD066613-06) for research, administrative, and computing support. The views expressed here do not represent those of NIH.

COLORADO.EDU

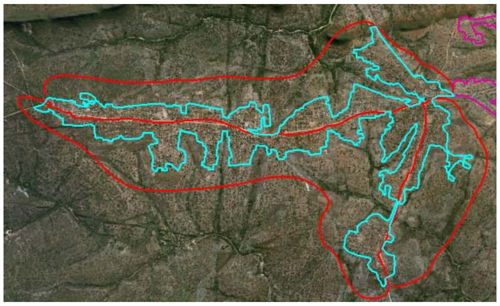
IBS
INSTITUTE OF BEHAVIORAL SCIENCE
UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Examples of three different clustering techniques:

1) clustering by village sections,



2) with the addition of roads as boundaries,



3) clusters based on distance from village edge.