# Movement patterns of farmers and forest workers from the Thailand-Myanmar border

*Edward Ho, Kailey Belcher, Leslie Vazquez Moreno, Marisa Passarella*
*Mentors: Daniel Parker, Vladimir Minin*

2022-08-19

## Introduction and Problem Statement

Human movement patterns play a significant role in infectious disease epidemiology, especially when infected individuals travel between areas of active transmission and disease-free areas. These movement patterns affect lower income nations the most as socio-economic instability and a lack of resources hinder disease control and increases susceptibility. Therefore, researching rural areas of low and middle income nations can provide insight on human components of epidemiological systems for many diseases, especially malaria. A recent study aimed to assess the feasibility of GPS loggers to empirically measure human travel patterns. This project uses the same data from that study of 53 participants to measure human movement patterns of villagers from the Thailand-Myanmar border while also quantifying different patterns by age, gender, season, and geographical areas. The goal of the project revolves around two objectives: (1) determine an acceptable buffer size around an individual's household to adequately capture the geographical landscapes they most commonly travel to near their residence and find any potential patterns of where individuals travel relative to the landscape buffer around their home. And (2) build a model to predict the geographical landscape an individual spends the majority of their time over the course of a week by using age, gender, season, and the total time they spend at home as predictors.

## Background and related work

Human movement impacts infectious disease epidemiology as many diseases are native to certain geographical locations and features. Malaria, for instance, occurs mainly in poor tropical or subtropical areas of the world, especially in Africa and Southeast Asia. Socio-economic instability and limited resources in these poorer areas hinder malaria control and increase the susceptibility of acquiring the disease.[1] Access to treatment also contributes to the spread and severity of disease. In fact, from 2010 to 2018, the number of Malaria cases in Myanmar not treated doubled to one in five.[2] However, if researchers can further understand citizens' movement patterns, they can implement policies or techniques to control and eliminate disease.

There are many methods that have been used to record human movement patterns either empirically or through personal observations. Personal observations include interviews, surveys,

and travel diaries; however, issues may arise if people choose to limit what they report (i.e. doing something they do not want others to know about). Empirical measures include mobile phone data and GPS loggers. A study in 2014 assessed the difference between travel data from surveys and mobile phones and found that surveys provide more demographic information while phones supply better spatio-temporal descriptions.[3] Another study successfully used GPS loggers to measure human mobility patterns in developing countries to then construct a model for disease transmission.[4] With constant technological advances, GPS loggers pave the way for spatio-temporal data collection and analysis, and this project aims to use GPS logger data to model the land types of where individuals spend most of their time.

# Data and Exploratory Data Analysis

In the study, there were 53 participants from the Thailand-Myanmar border. The study size was intentional as this was an exploratory pilot study. Each participant received a GPS logger to carry with them from March 2017 to February 2018. The participants were recruited from two clinics and then put into groups according to which village they resided near said clinics. Participants were adults from the Karen or Burmese ethnic groups that stated they could carry the GPS logger and were able to walk outside of village boundaries. The loggers were programmed to take a reading every 30 minutes and lay dormant if no movement was detected for longer than one hour. Devices were also programmed to take readings at one-minute intervals if it was moving more than 15km/hour which accounts for traveling by vehicle rather than walking. Further information on data collection and recording can be found at https://wellcomeopenresearch.org/articles/6-148.[5]

The original dataset had 314,482 observations with 11 variables, the most relevant being date, time, latitude, longitude, and distance. Another dataset included the demographic list of participants: their logger code, name, age, and sex. The age and sex variables were taken from this set and added to the original. There were 44 males and only 9 females in the study; most females ranged from 29 to 50 years old while males ranged from 21 to 39 years old. Figure 1 shows a histogram of the male and female observations. Since the project is interested in the seasons, the seasons were coded as “cool and dry season” for mid-October to mid-March, “hot and dry season” for mid-March to mid-May, and “rainy season” for mid-May to mid-October.[5] Figure 2 represents the percentage of time spent at home throughout the year based on season and gender. Other notable variables include the location of where they slept, the distance they traveled from that location throughout the day, how they allocated their time throughout the day, and the geographical description of their current location and where they slept that night. The predictor variable was coded as “geo_max” which represents the environmental land type that a person spends most of their time at during one week. Assumptions were made when creating these variables. Firstly, it was assumed that if their trackers were off, the time passed counted as them staying home. Secondly, their home/sleep location was calculated with respect to the last tracked location of the day.

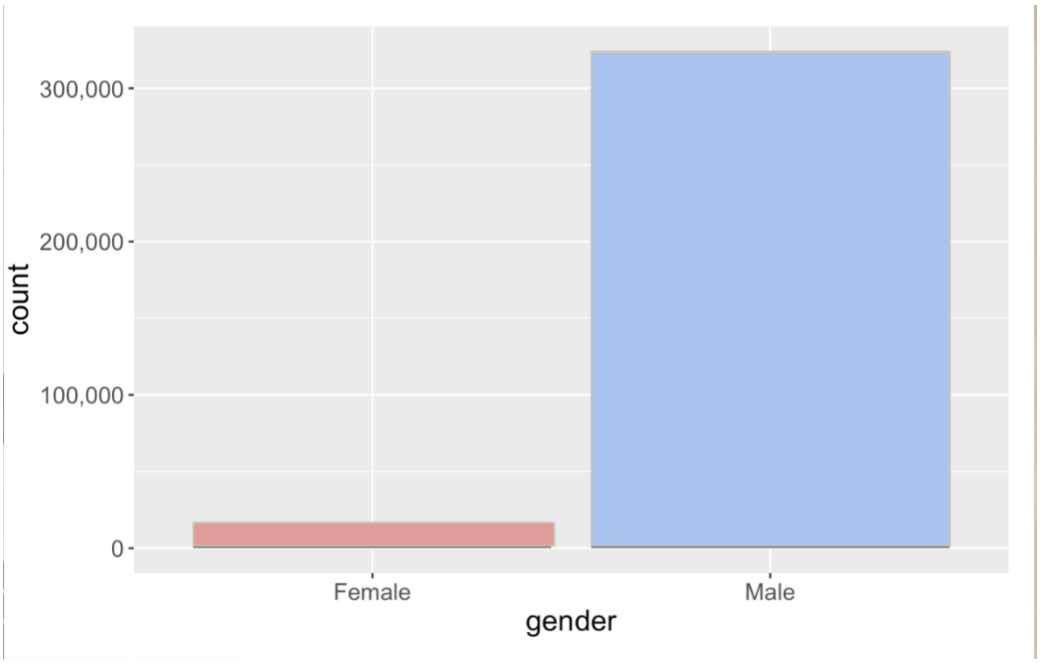


Figure 1: Observations of males versus females in the study.

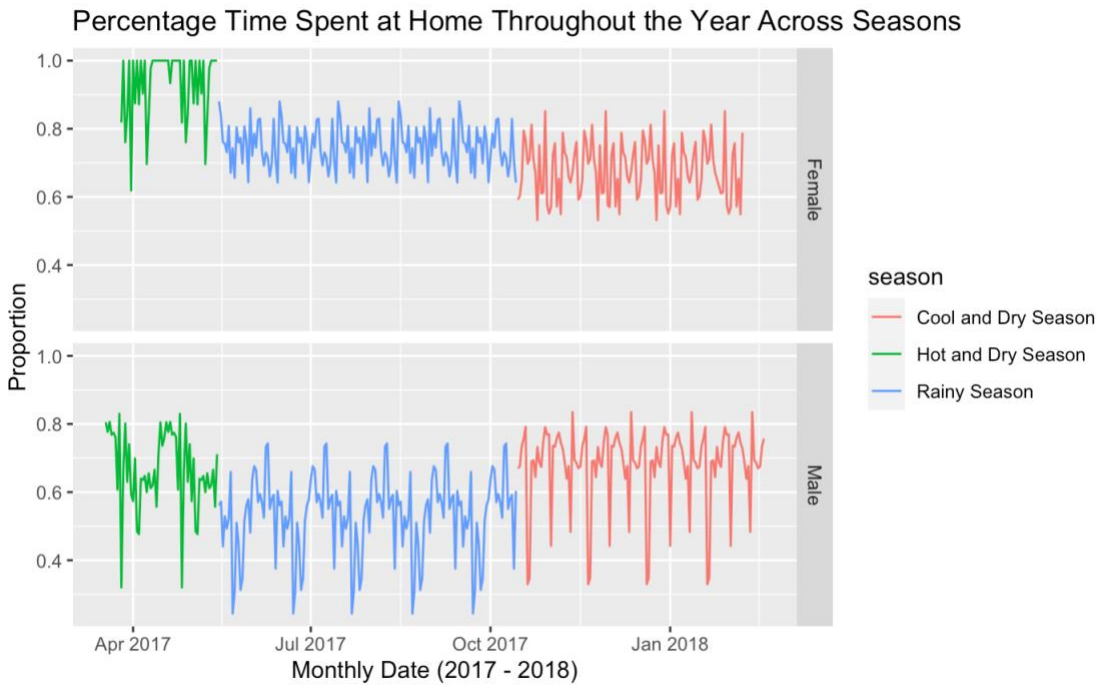


Figure 2: Percentage of time spent at home throughout the year for males and females based on seasons.

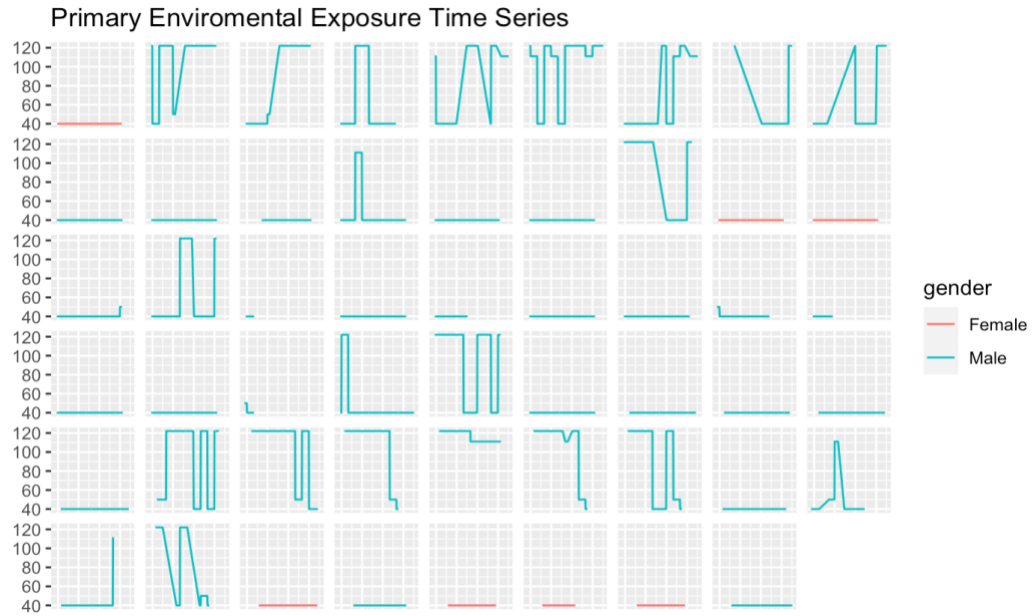


Figure 3: Where participants spent the majority of their time weekly over the course of the entire study. The numbers on the y-axis represent the different geographical land types from cultivated vegetation, urban areas, and forests.

The project was interested in classifying the land surfaces that the participants were traveling to; these surfaces are being called environmental exposures. To observe the types of land people traveled to, the Copernicus Global Land Service Land Cover raster dataset was used.[6] This dataset uses pixelated data associated with geographical locations to represent the land surface of that location.[6] This dataset includes map codes for the land types, so the participant's latitude and longitude coordinates were matched with the correlated land cover map codes. This then was used to map the coordinates onto the land surfaces that the participants traveled to throughout the study. The areas of interest where travelers visited the most can be found in table 1 below with the matching key. These codes are used when classifying the land cover classes in the model.

| Map Codes | Land Cover Class |
|---|---|
| 112-116 | Open forest |
| 121-124 | Closed forest |
| 20 | Shrubs |
| 30 | Herbaceous vegetation |
| 90 | Herbaceous wetland |
| 40-47 | Cultivated and managed vegetation/ agriculture (cropland) |

| | |
|---|---|
| 50 | Urban/built up |
| 80 | Permanent water bodies |
| 200 | Open sea |

Table 1: Map codes with associated land cover class.

This data is essential to look at the human movement patterns because it classifies the coordinates based on land types which can be used to infer a person's environmental exposures. From this, one can determine the percentage of time that a person spends in each land surface type. Plotting all of the GPS coordinates of each participant over the course of the study gives a visual representation of the environment exposures on the Thailand-Myanmar border, as seen in Figure 4.

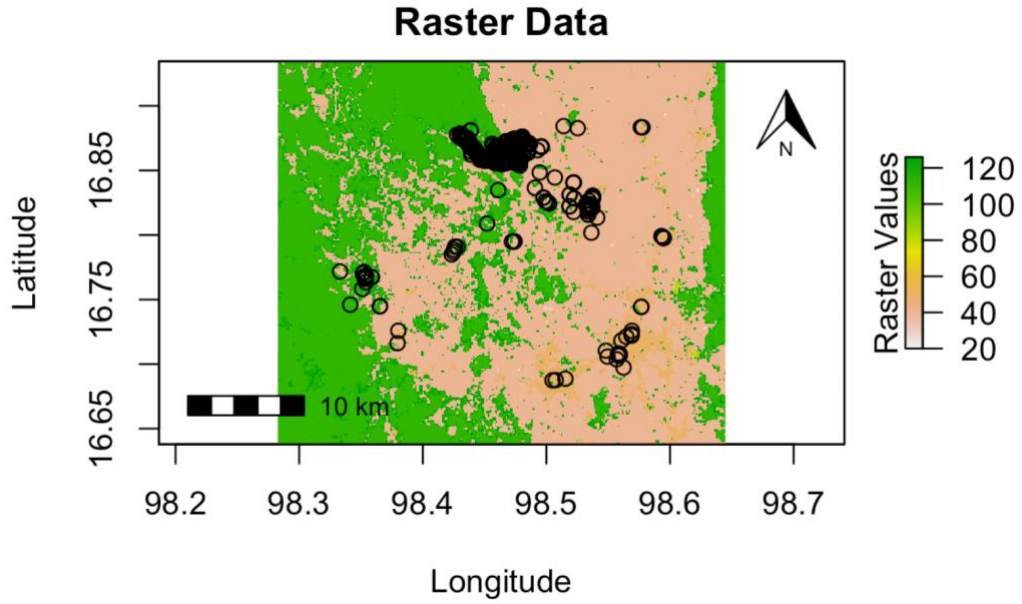


Figure 4: The raster plot of one participant which shows they have traveled within cultivated land, forest, and urban areas.

# Methods

### Raster and Buffer Sizes

The Copernicus Global Land Service Land Cover[6] was used to determine the types of land individual's traveled to throughout the study. Since home locations were not provided, these were considered to be the last GPS tracked location of each day. This takes into the account that a person may have one or more sleeping locations. For example, farm workers could stay in a secondary home location during the cultivating season. From this, four different sized buffer zones were created around these sleeping locations: 250m, 500m, 1000m, and 2000m. The 250m buffer size was found by taking the median distance of all the maximum travel distances from a

person's sleeping location and then taking all the values from 0-50% of those distances and calculating the mean. This method ensured that among all participants, there was a likely chance they were residing in their home residence and after exiting the 250m zone, they were considered to be traveling to other areas. The other buffer sizes were made arbitrarily to test the validity of the 250m buffer. Within these buffers, the proportion of land types were observed and then placed in the model to see if buffer size affected the predictor variable.

**Modeling**

Multinomial logistic regression was used to model the outcome since it is categorical and there were more than two categories (geographical locations). The predictors incorporated into the model were age, season, and each proportion of the individual's geographic location within the buffer. Originally, gender was used as a predictor, but due to a lack of diverse observations, gender was eliminated from the model. Multinomial logistic regression uses "stacked" logistic regression where the outcome represents the log odds of the probability of the geographic location over the probability of the base geographical location. The base geographical location is the first geographical location which is 40, cultivated vegetation and agriculture. Twelve models were constructed and were divided into either fixed effects, mixed model with effect on season, or mixed model with random effect on the individual. The mixed effect differs from the fixed effect since the mixed begins the regression at an intercept specific to the random effect. A Categorical Regression book was used to reference model assumptions.[7] These assumptions include independence of irrelevant alternatives (IIA), a categorical outcome, the log-odds of the outcome and independent variable have a linear relationship, errors are independent, and no severe multicollinearity. The assumptions were met and are further explained in the results section.

# Results

AIC and BIC were used to measure the model. The model with random effects on the individual was the better model because it had the lowest AIC and BIC compared to the others.
The fixed models and the mixed model with random effect on season did not vary much in AIC and BIC and neither did the buffer sizes within each of the models, although 250 meters had the least AIC and BIC among them all.

As for the results of the best model, many of the p-values were not statistically significant, all yielding a mildly to extremely high p-value between 0.3-0.8. However, the significant predictors were all season when it came to open and closed forest areas. This most likely indicates that most people have a seasonal reason to visit the areas of open and closed forest, whereas urban and agricultural areas are the primary areas of focus regardless of season. One interesting result was that many people in urban areas were extremely likely to stay within urban areas over the course of the study as their most visited location. Regarding the buffer size, the result from the model specific to the data provided showed that environmental exposure proportions overall do not

greatly affect the movement of individuals outside of an urban environment. Finally, regarding the mixed effects model, it was seen that open and closed forests were statistically significant effects, indicating that the people that went to open/closed forests were the ones consistently returned while urban areas were inclusive of everybody.

Each of the assumptions were met. The IIA assumption is met because all of the possible geographic locations are incorporated in the model and adding any additional locations would not change the relative likelihood of the geographic locations. The log-odds of the outcome and independent variable have a linear relationship, and errors are independent, as seen in the residual fit in Figure 5.

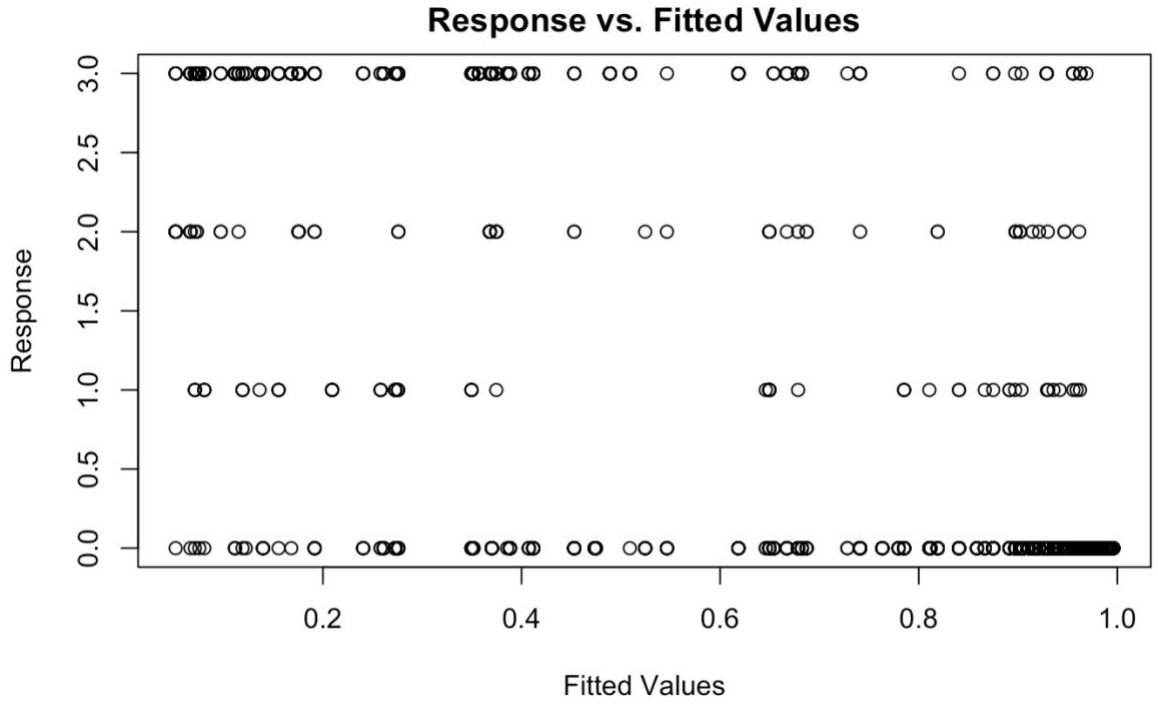


Figure 5: Response vs fitted values plot. The fitted values are the predicted percentages of where they visit and response is 0, 1, 2, 3, which respectively represent agricultural, urban, open, or closed forest.

## Discussion and Conclusion

Overall, it was shown that home buffers generally are not the best teller of where an individual will spend their time. Season was the best predictor of where someone would spend the majority of their time. This was expected as season impacts working conditions, especially on a farm. One thing to note is that the primary reason the model may have had interesting fit was that the data contained primarily agricultural land values which allows the model to easily predict those areas and limits its ability to discern any differences between the others. The project had quite a few limitations. For instance, home GPS locations were not collected and instead were calculated

based on sleeping locations to then create the buffer which could not be fully accurate. Additionally, GPS batteries only last 1-2 months, so some data could have been missed in between the loggers being exchanged. The study had a small sample size since there were only two clinics, so future studies should have more participants to create a better model.

In the future, similar research should conduct a more inclusive study since this one did not have many females. There are many different ways to create buffer sizes, so future work can also be done to explore the best technique for developing adequate buffer sizes. Lastly, if GPS logger data can be combined with infectious disease data, such as malaria cases, a future study can model how locations are related to malaria transmission. Ideally, this could ultimately aid in disease control and the elimination of the disease in Southeast Asia as a whole.

## References

1. Centers for Disease Control and Prevention. (2021, December 16). *Malaria's Impact Worldwide*. Centers for Disease Control and Prevention. Retrieved August 11, 2022, from https://www.cdc.gov/malaria/malaria_worldwide/impact.html
2. OECD iLibrary. (n.d.). *Malaria*. Health at a Glance: Asia/Pacific 2020 : Measuring Progress Towards Universal Health Coverage . Retrieved August 11, 2022, from https://www.oecd-ilibrary.org/sites/8d22b645-en/index.html?itemId=%2Fcontent%2Fcomponent%2F8d22b645-en
3. Wesolowski, A., Stresman, G., Eagle, N. *et al.* Quantifying travel behavior for infectious disease research: a comparison of data from surveys and mobile phones. *Sci Rep* 4, 5678 (2014). https://doi.org/10.1038/srep05678
4. Vazquez-Prokopec GM, Bisanzio D, Stoddard ST, *et al.*: Using GPS Technology to Quantify Human Mobility, Dynamic Contacts and Infectious Disease Dynamics in a Resource-Poor Urban Environment. Colizza V, editor. *PLoS One.* 2013; **8**(4): e58802.
5. Tun STT, Min MC, Aguas R *et al.* Human movement patterns of farmers and forest workers from the Thailand-Myanmar border [version 1; peer review: 1 approved with reservations]. *Wellcome Open Res* 2021, **6**:148 (https://doi.org/10.12688/wellcomeopenres.16784.1)
6. Marcel Buchhorn, Bruno Smets, Luc Bertels, Bert De Roo, Myroslava Lesiv, Nandin-Erdene Tsendbazar, Martin Herold, & Steffen Fritz. (2020). Copernicus Global Land Service: Land Cover 100m: collection 3: epoch 2017: Globe (V3.0.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3518036
7. Werth, Rose. "Categorical Regression in Stata and R." *15 Multinomial Logit Regression (R)*, https://bookdown.org/sarahwerth2024/CategoricalBook/multinomial-logit-regression-r.html.