Harriet Gislam, 1st year PhD student, School of environment and technology, University of Brighton

Email: H.gislam1@uni.brighton.ac.uk Supervised by: Dr Niall Burnside, Dr Matthew, Dr Kebede Deribe and Prof Gail Davies

National Institute for Health Research

University of Brighton

LINKING SOILS AND HUMAN HEALTH: GEOSPATIAL ANALYSIS OF PODOCONIOSIS OCCURRENCE AND CAUSE IN CAMEROON

Project aim: The aim of this project is to investigate in detail the chemical and mineralogical soil variables related to podoconiosis prevalence, and to develop a fine-scale risk model of podoconiosis for Cameroon. A high resolution risk map will be developed to assist identification of locations for targeted initiatives seeking to eliminate this neglected tropical disease.

Objective 1: To investigate soil chemical and mineralogical data to identify specific soil mineralogical correlates with contemporary podoconiosis prevalence data, and proximities, using geo-statistical analysis.

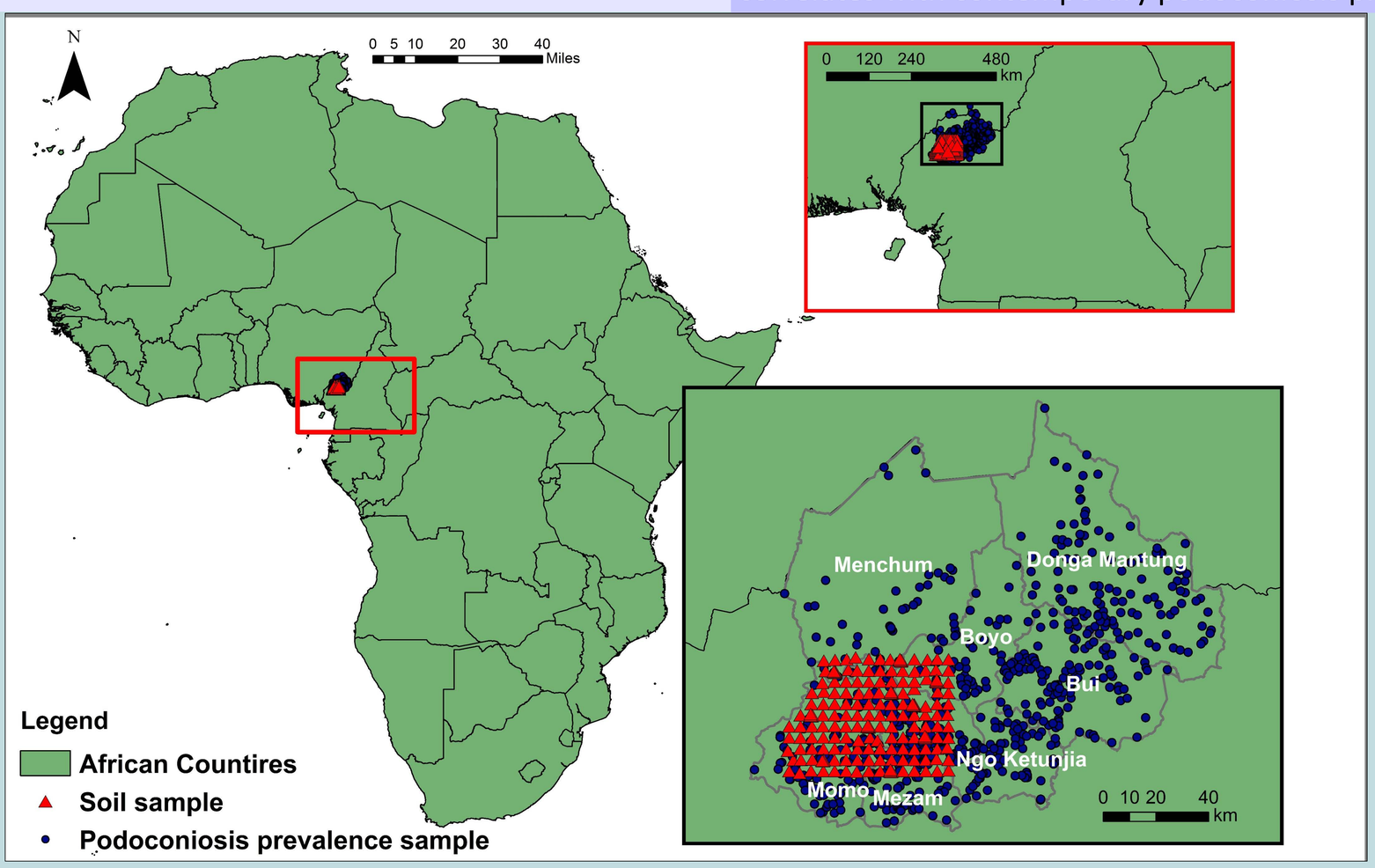

Figure 1: Map of Africa, with two inset maps of Cameroon. Inset map shows location of soil and podoconiosis prevalence sample locations.

Data collection

Soil sampling data: Soil sampling data was collected at 152 sampling sites equally spaced along a grid, samples were taken in the centre of each grid square (Le Blond, unpublished).

There were 10 grid squares that contained 5 extra sampling sites, to capture the soil variability at a greater resolution. In the 10 grids with the multiple samples the sampling technique seen in figure 2 was utilised. Soils were tested using ICP-MS and XRD.

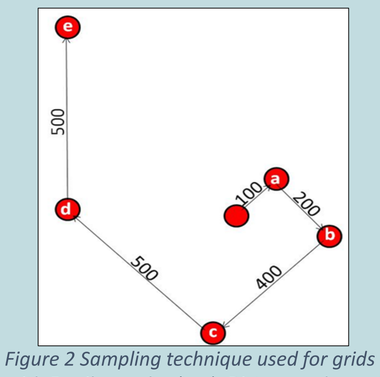

Figure 2 Sampling technique used for grids with 5 soil samples (a-e), Distances showed in m (Le Blond, unpublished).

Podoconiosis prevalence data: The podoconiosis prevalence data was collected by trained Community Health Implementers (CHI) at community level representing a total of 672 samples. At each community, GPS coordinates, eligible population and number of podoconiosis cases was recorded. Prevalence data was adjusted due to recognised inaccuracies of the CHI's in identifying cases of podoconiosis (Wanji et al., 2016). The 168 prevalence data points, most proximal to soil samples were utilised.

Methodology

Interpolation: Soil chemical and mineralogical data was interpolated, comparing methods of inverse distance weighting, ordinary kriging, empirical Bayesian kriging and universal kriging. Interpolation surface selection was based on prediction error statistics: mean error (Eq.1), root mean squared error (Eq.2), average kriging standard error (Eq.3), mean standardised prediction error (Eq.4) and root mean square standardised prediction error (Eq.5).

$$ME = \frac{1}{N}\sum_{i=N}^{N}\{Z(x_i) - \hat{z}(x_i)\} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=N}^{N}\{z(x_i) - \hat{z}(x_i)\}^2} \quad (2)$$

$$ASE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sigma^2\,(xi)} \quad (3)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\frac{ME}{\sigma^2(x_i)} \quad (4)$$

$$RMSSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{ME}{\sigma^2(x_i)}\right)} \quad (5)$$

Figure 3: list of prediction error statistics.

N the number of values in the data set, $\hat{z}(x_i)$ the predicted value, $z(x_i)$ the observed value and σ^2 is the kriging variance for location x_i (Eq.1-2) are applicable to all interpolation techniques whereas (Eq.3-5) are only applicable for kriging as they require kriging variance. The ideal model will have mean error equal to 0, the smallest-root-mean-square error, the average standard error nearest the root mean square error and the standardised root mean square error closet to 1 (Johnston et al., 2001). These prediction statistics were compared for each model.

Univariate analysis

There has been no studies in Cameroon which investigate daily activity areas. A 3km buffer was created, to represent daily activity areas. A 3km distance was chosen, due to a study in Uganda which determined village activities distance at 3km from village centroid (Wardrop et al., 2012). Using the zonal statistics as table function in ArcMap, soil variables values were extracted and a mean value was calculated from the 3km buffers.

Spearman rho correlation and Kruskal-Wallis was carried out between podoconiosis prevalence and soil variables. These tests were used to support the selection of variables for the multivariate model. Soil variables with a significant correlation of r >0.7 between each other were not be selected for the model, due to the effect of multicollinearity (Dormann et al., 2013). If variables were correlated r >0.7 the one with the strongest correlation to podoconiosis prevalence or statistically significant Kruskal-Wallis statistic were selected.

Multivariate analysis

Hurdle model

The hurdle model is utilised for its ability to handle excess zero values, which are found in the prevalence data. Hurdle models are two-part models that model the zeros separately from the positive counts (Mullahy, J. 1986). The first part is the hurdle component which models the zero counts, usually a binomial model. The second part the count model can be a truncated poisson or negative binomial. The count part of the model is only employed if the hurdle (prevalence >0) is exceeded. R was utilised for creating the model using the package pscl and the command hurdle(). AIC was used to select model components and choose the soil variables for both parts of the model out of those selected from the univariate analysis. Only statistically significant variables with a p-value <0.05 were chosen for the final model using stepwise backwards elimination.

Results

By comparing the AIC value, the model selected was binomial with logit link for the binary part of the model and truncated poisson with log link for the count part. Table 1 shows the soil variables selected for the final model through stepwise backwards elimination. For both parts of the model the odds ratio and risk ratio are all close to 1, so although significant the results are not conclusive.

The next steps will be to investigate other multivariate models such as zero-inflated models and will consider more advanced machine learning modelling. Different buffer sizes around prevalence data points will also be explored to determine if this has any significant effect on model outcomes.

Table 1: Result from the hurdle model, the zero binary part of the model (binomial with logit link) and the count part (truncated poisson with log link).

Count (truncated poisson with log link)	Coefficent	SE	P-value	RR	LCL	UCL
Intercept	1.548642	0.962955	0.10779	4.71	0.71269	31.0624
Ba	-0.008191	0.003003	0.00638	0.99	0.98602	0.9977
Be	-0.027585	0.005903	3E-06	0.97	0.9616	0.98411
Cd	-0.007087	0.002394	0.00307	0.99	0.98829	0.99761
Fe	0.014298	0.003973	0.00032	1.01	1.00653	1.02233
Hydrobiotite	-0.030835	0.004541	1.1E-11	0.97	0.96104	0.9783
Mica	0.026441	0.006311	2.8E-05	1.03	1.01417	1.03957
LOI	-0.018875	0.003668	2.7E-07	0.98	0.97427	0.98838
V	0.029567	0.005216	1.4E-08	1.03	1.01953	1.04059
Zero (binomial with logit link)	**Coefficent**	**SE**	**P-value**	**OR**	**LCL**	**UCL**
Intercept	-5.020789	1.590782	0.0016	0.006	0.00021	0.42391
Al	0.013584	0.006459	0.03545	1.01	1.00276	1.03059
As	-0.018735	0.007985	0.01896	0.98	0.95989	0.99764
Be	0.023391	0.007477	0.00176	1.02	1.01268	1.04743
Fe	-0.012232	0.005298	0.02095	0.99	0.96928	0.99562
Quartz	0.010946	0.005292	0.03861	1.01	1.00265	1.02978
W	0.020784	0.008214	0.01139	1.02	1.00262	1.03916

(SE)= standard error, (RR)= risk ratio, (OR)= odds ratio, (LCL)= lower confidence limit, (UCL) = upper confidence limit.

Conclusion

The hurdle model identified several variables which are statistically significant, to podoconiosis occurrence and prevalence. Although the results are by no means conclusive, this initial investigation highlights variables which have the potential of predicting podoconiosis prevalence. The exploration of more advanced modelling on the data will further the understanding of the potential to predict podoconiosis occurrence.

Further project objectives

2: To explore the potential of multispectral and hyperspectral satellite data to map and predict fine-scale soil variability with the use of machine learning algorithms.

3: To develop a novel spatial modelling technique, using multi and hyper-spectral data, to produce and validate a fine-scale risk map of podoconiosis prevalence in Ethiopia and Cameroon

References

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J. and Münkemüller, T., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), pp.27-46.

Johnston, K., Ver Hoef, J.M., Krivoruchko, K. and Lucas, N., 2001. *Using ArcGIS geostatistical analyst* (Vol. 380). Redlands: Esri.

Mullahy, J., 1986. Specification and testing of some modified count data models. *Journal of econometrics*, *33*(3), pp.341-365.

Wanji, S., Kenge-ouafo, J. A., Datchoua-Poutcheu, F. R., Njouendou, A. J., Tayong, D. B., Sofeufeugaing, D. D., Amvongo-Adjia, N., Fovennso, B. A., Longang-Tchounkeu, Y. F. & Tekola-Ayele, F. 2016. Detecting and staging podoconiosis cases in North West Cameroon: positive predictive value of clinical screening of patients by community health workers and researchers. *BMC public health,* 16**,** 997.

Wardrop, N. A., Fevre, E. M., Atkinson, P. M., Kakembo, A. S. & Welburn, S. C. 2012. An exploratory GIS-based method to identify and characterise landscapes with an elevated epidemiological risk of Rhodesian human African trypanosomiasis. BMC infectious diseases, 12, 316.