**Read Me**

This document contains instructions for replicating the results for the article “Comparing Random Forest with Logistic Regression for Predicting Class-imbalanced Civil War Onset Data”. Replication materials include replication R code, CSV files containing replication data, a .pdf file specifying the variables used in the analysis taken from Hegre and Sambanis 2006 and this read me file.

R Code:

There is one R file containing all the code needed to replicate the results for the article “Comparing Random Forest with Logistic Regression for Predicting Class-imbalanced Civil War Onset Data”. The code is commented such that all directions for replication are easily recognizable. The file name for the R code is Comparing Random Forest with Logistic Regression R Code.R.

Data:

There are two data files. The first, SambanisImp.csv contains the fully imputed dataset taken from Hegre and Sambanis (2006). This first dataset was used to explore the predictive accuracy of random forests compared to logistic regression. Figures 1,2, and 4 were constructed from this data. The second data file, Amelia.Imp3.csv was used to explore causal mechanisms related to civil war onset. Amelia.Imp3 is a smaller dataset where only variables theorized to be most relevant to the onset of civil war were imputed. Variables removed from the Amelia dataset were chosen based on theory and regression models included in Fearon and Laitin (2003), Collier and Hoeffler (2004), ad Hegre and Sambanis (2006).

Codebook

There is a codebook included titled Sambanis Codebook. It contains names a descriptions of all the variables collected and used in Hegre and Sambanis (2006) and hence our data.