Overview of epidemiar Package

Dawn Nekorchuk, Michael Wimberly, and EPIDEMIA Team Members
Department of Geography and Environmental Sustainability, University of Oklahoma
dawn.nekorchuk@ou.edu; mcwimberly@ou.edu

Updated April 19, 2019

Contents

1 Introduction

The Epidemic Prognosis Incorporating Disease and Environmental Monitoring for Integrated Assessment (EPIDEMIA) Forecasting System is a set of tools coded in free, open-access software, that integrate surveillance and environmental data to model and create short-term forecasts for environmentally-mediated diseases.

This R package, `epidemiar`, is the cornerstone of the forecasting system, and is designed to be used to model and forecast a wide range of environmentally-mediated diseases.

1.1 Issues/Background

1. Public health monitoring of environmentally-mediated diseases can benefit from incorporating information on related environmental factors. This melding of data can improve the ability to detect early indication of outbreaks, allowing for more efficient and proactive public health interventions.
2. Originally, the EPIDEMIA project integrated local malaria surveillance data and remotely-sensed environmental predictors to produce operational malaria forecasts for the Amhara region of Ethiopia. Our local public health partners expressed interest in being self-sufficient in creating the weekly reports themselves.

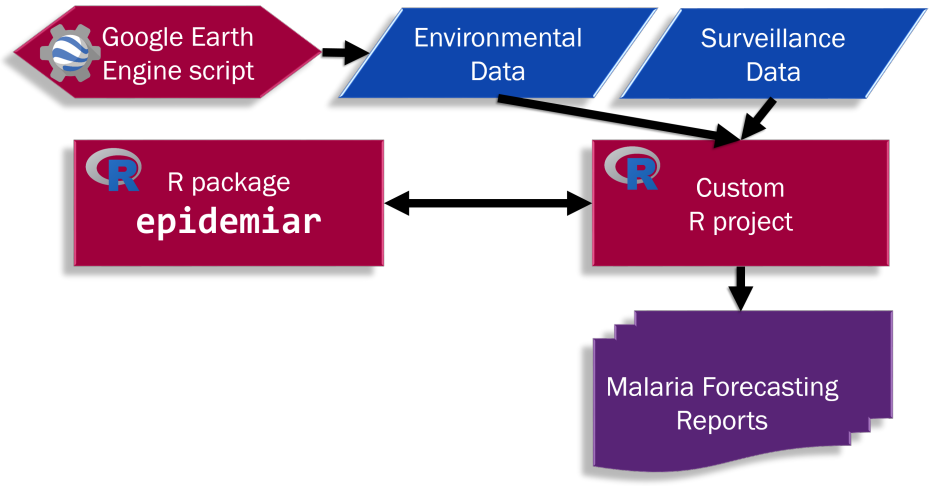
1.1.1 Our Solution

The EPIDEMIA modeling of disease transmission for early detection and early warning evaluation had been done in R. We developed the "epidemiar" R package to provide a generalized set of functions for disease forecasting.

In addition, we designed workflows and wrote customized code for our Ethiopian colleagues, including a Google Earth Engine script to capture the necessary summaries of the environmental variables. The output of the modeling and forecasting is fed into formatting documents to create distributable reports with maps and graphs of the results. A version of this code is being released as `epidemiar-demo`: https://github.com/EcoGRAPH/epidemiar-demo.

1.2 System Components

The full system can be thought of have 3 (three) main parts:



1. R package `epidemiar`: This package - a library of flexible functions for modeling and forecasting
2. Google Earth Engine script: Script to gather the environmental data summaries
3. Custom R Project: Contains the surveillance and environmental data, user parameters on the model and outbreak detection algorithm, and script to produce a finalized report.

This package can be used for modelling and forecasting for a variety of environmentally-mediated disease. For example GEE scripts and R project, see the `epidemiar-demo` repository at https://github.com/EcoGRAPH/epidemiar-demo.

The main requirements for using this package are:

- surveillance / disease case counts per week per geographic group
- daily environmental data per geographic group with enough lead time for lagged effects (user set)
- pre-identified model: which environmental covariates to include, any clustering of geographic groups.

2 Modeling Overview

The model is based on a general additive model (GAM) regression of multiple factors, including the geographic group, long terms trends, seasonality, lagged environmental drivers and clustering of geographic groups.

$$
\begin{aligned}
log(cases) \sim\ & geo + bs_1 * geo + bs_2 * geo + bs_3 * geo + bs_4 * geo + bs_5 * geo + \\
& s(doy, bs = "cc", by = geo) + \\
& (env_1 sum_1 * cl + env_1 sum_2 * cl + env_1 sum_3 * cl + \\
& \quad env_1 sum_4 * cl + env_1 sum_5 * cl) \ + \cdots \\
& (env_n sum_1 * cl + env_n sum_2 * cl + env_n sum_3 * cl + \\
& \quad env_n sum_4 * cl + env_n sum_5 * cl)
\end{aligned}
$$

where *geo* is the geographic group, $bs_1 \ldots\ bs_5$ are modified basis functions, *doy* is the day of the year, *env* are the environmental variables (1, 2 ... n) and the 5 summary (*sum*) statistics from the lagged basis functions, and *cl* is the cluster identification of that geographic group. The regression is done with `family=poisson()` for a log link function to the case count. See the following sections for more details.

2.1 Geographic group, long term trends, and seasonality

Each geographic group, *geo*, identified in the `groupfield` column is included as a factor.

To capture any long term trends per geographic group, *geo* is multipled by each of 5 modified basis splines: $bs_1 * geo + bs_2 * geo + bs_3 * geo + bs_4 * geo + bs_5 * geo$. The modified basis splines are created within the function as follows:

1. First, `splines::bs()` is used to create basis splines over the range of observations with degree 6.
2. To reduce the edge effects of using splines, the following modifications are performed:
    - the last basis spline function is reverse, and
    - the second to last basis spline function is removed.

To account for seasonality in each geographic group, a cyclical cubic regression spline smooth is added based on day of year per geographic group: $s(doy, bs = "cc", by = geo)$

2.2 Environmental Variables

The rates of environmentally-mediated infectious diseases can be influenced by the environmental factors via a range of potential mechanisms, e.g. affecting the abundance and life cycle of disease vectors. The influences on disease generally lags behind the changes in the environmental covariates.

In our modeling, the **anomalies** of the environmental covariates are used as factors. We are looking at the influence of deviation from normal in the environmental factors to help explain deviations from normal in the human cases. The variables were transformed by taking the residuals from a GAM with geographic unit and cyclical cubic regression spline on day of year per geographic group: $geo + s(doy, bs = "cc", by = geo)$

In the modeling controls, the user selects the maximum number of days in the past (lag length, l) to consider interactions. Each geographic group and week is associated with environmental anomaly values on the day the week began, up to the lag length, l, so that each group-week has a l-day history. A distributed lag basis is created with the natural cubic splines function (ns, splines library), including intercept, with knots at 25%, 50%, and 75% of the lag length. The 5 basis functions that result are multiplied by each group's history, so that there are just 5 summary statistics, instead of l, for every combination of group, week, and environmental anomaly covariate.

2.3 Clusters

The relationship between environmental drivers and the case burden of the environmentally-mediated disease can vary with geographically, due to ecological, social or other geographic factors. This potential spatial non-stationarity could be handled in a number of ways.

If you were working with areas not likely or shown not to have spatial non-stationarity between environmental covariates and disease rates, you could use a global model (all geographic groups in one cluster). However, if there are spatial variations in environmental influence, this could yield a poorer model fit.

On the other extreme, you could run separate models for each geographic group (each geographic group as its own cluster). However, especially with noisy data or short time-series, this could lead to overfitting.

We allow the user to identify their own clusters of geographic units. The clustering determination can be done prior however the user chooses - for example, global model, individual models, clustering by ecological zones, or by identifying similar temporal disease patterns.

3 Event Detection Overview

The central idea behind outbreak detection is to identify when the case volume exceeds a baseline threshold, and to use this information in a prospective (not retrospective) manner to identify epidemics in their early stages.

Currently, epidemiar supports the Farrington improved algorithm for event detection, using `surveillance::Farringtonflexib`

This family of methods developed by Farrington and later, Noufaily, have been implemented at several European infectious disease control centers. Farrington methods are based on quasi-Poisson regression and

can take advantage of historical information while accounting for seasonality, long-term trends, and previous outbreaks.

The Farrington improved method offer parameters to control various model settings, such as the number of time points to include in the historical window through a specified number of years, the number of knots in the splines to account for seasonality, and the number of weeks to exclude at the beginning of the evaluation period (for events that may already be in progress). However, this method does generally require several years of historical data.

Alerts are generated by the Farrington algorithm run over the entire time length of the report on the number of cases, observed or future forecast (optionally adjusted for population).

Early Detection alerts are alerts that are triggered during the early detection period, a user set number of week of the most recently known epidemiological data (case counts).

Early Warning alerts are alerts that are triggered in the future forecast estimates (early warning period). These early warning alerts indicate that the environmental conditions are favorable (or unfavorable) for abnormally high case counts, based on past trends.

Alerts per week per geographic group are recorded. As the algorithm runs over the entire length of the report, historical alerts (weeks included in the report that are prior to the early detection period) are also marked.

Alert summaries are also created for the early detection and early warning periods (separately). "High" level indicates two or more weeks in this period had incidences greater than the alert threshold, "Medium" means that one week was in alert status, and "Low" means no weeks had alerts.

4 References

Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. J R Stat Soc Ser A Stat Soc. 1996;159:547–63.

Höhle M. surveillance: An R package for the monitoring of infectious diseases. Comput Stat. 2007;22:571–82.

Hulth A, Andrews N, Ethelberg S, Dreesman J, Faensen D, van Pelt W, et al. Practical usage of computer-supported outbreak detection in five European countries. Eurosurveillance. 2010;15:19658.

Merkord CL, Liu Y, Mihretie A, Gebrehiwot T, Awoke W, Bayabil E, et al. Integrating malaria surveillance with climate data for outbreak detection and forecasting: the EPIDEMIA system. Malar J. 2017;16.

Noufaily A, Enki DG, Farrington P, Garthwaite P, Andrews N, Charlett A. An improved algorithm for outbreak detection in multiple surveillance systems. Stat Med. 2013;32:1206–22.